Hibrid CNN hálózat egyetlen kép alapú mélység becsléséhez: egy esettanulmány *

Károly Harsányi[1], Attila Kiss[1], András Majdik[1], and Tamas Sziranyi[1 2]

[1] Machine Perception Research Laboratory, MTA SZTAKI, Budapest, Hungary
[2] Faculty of Transportation Engineering and Vehicle Engineering, BME, Hungary
{harsanyika, attila.kiss, majdik, sziranyi}@sztaki.hu

Kivonat A háromdimenziós képfeldolgozásnak számos alkalmazásban jelentős szerepe van az autonóm vezetés, a robotika és a folyamatos, valós idejü pályakövetés területén. Ennek egyik fontos eleme a gyorsabb és megbízhatóbb algoritmusok tervezése és fejlesztése, amelyek képesek RGB-képek alapján a mélység becslésére. Célunk tehát egy olyan algoritmus kidolgozása, amely egy képen a lehető legpontosabban meg tudja becsülni minden pixel mélységét egyetlen nézetből. Tanulmányunkban a mélytanulás lehetőségeit kihasználva a mélységi előrejelzés mély konvolúciós neurális hálózatok alkalmazásával történik. A jobb eredményeket eléréséhez a javasolt architektúra új mellékoldali kapcsolatokat tartalmaz a kódolási és dekódolási ágak között.

Keywords: mélység becslés · mély konvolúciós neurális hálózat · CNN

1. Introduction

Predicting the depth of the elements of a scene has been an interesting challenge since the foundation of computer vision. It is a well known and well studied hot topic in the field of computer science. Among several techniques, some of the most known ones are using some special information from the images, for example, variations in illumination [29,25], or focus [4,1]. Another popular method in this category is the so-called Structure-from-Motion technique [27,20]. However, it is extremely difficult to gather any special information from a single image without additional knowledge about the environment. Nevertheless, finding better solutions to this problem is vital, considering that accurate depth information improves results on human pose estimation [26,16], recognition [17,5], reconstruction [21,13] or semantic segmentation [2,11].

* The research was supported by the Hungarian Scientific Research Fund (No. NKFIH OTKA K-120499 and KH-126513) and the BME- Artificial Intelligence FIKP grant of EMMI (BME FIKP-MI/FM). The paper was originally published at: Harsányi, K., Kiss, A., Majdik, A., Szirányi, T. A hybrid CNN approach for single image depth estimation: A case study. In K. Choros, M. Kopel, E. Kukla, A. Sieminski (Eds.), Multimedia and Network Information Systems - Proceedings of the 11th International Conference MISSI 2018 (pp. 372-381). (Advances in Intelligent Systems and Computing; Vol. 833). Springer Verlag. DOI: 10.1007/978-3-319-98678-4_38

Before neural network approaches became popular in this field, other approaches appeared based on Conditional Random Fields [23,8] (taking superpixels into consideration) and Markov Random Fields (MRF) [10,24]. Later, a very effective tool came into focus. With the help of Convolutional Neural Networks (CNNs) we became able to learn an implicit relation between depth and color pixels. Combining CNNs with CRF-based regularization via structured deep learning is a promising direction in the research [12,9].

In this paper, we introduce a fully convolutional neural network which combines the advantages of deep residual nets [6] and U-net architectures [18]. We examine the performance of this hybrid network on the NYU depth v2 dataset [21], and demonstrate that it is possible to achieve state-of-the-art accuracy without significantly increasing the depth and the number of variables. This is achieved by augmenting the CNN from the work of Laina et al. [9] with U-net-like connections between the encoding and decoding parts. This kind of augmentation is getting more and more common in the engineering society. We applied this successfully to a new domain.

2. Related Work

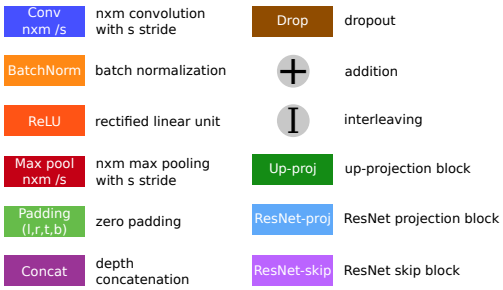
In the introduction, we referred some classical methods on depth estimation so, in this section, we try to focus on the latest results related to our work. Outstanding results and the possibility of new methodologies drove research towards the application of CNNs for the aforementioned problem. Classical networks, originally developed for a different task, were studied in this topic. These networks like AlexNet [7] and VGG [22] earned fame by their high level of success at the ImageNet Large Scale Visual Recognition Challenge [19]. Later, a two-scale architecture approach has appeared by Eigen et al. [3]. One year later, their work was extended to a three-scale architecture for further refinement in [2].

We have to mention two special network architectures in order to understand the motivation behind our approach. The next breakthrough in the field of neural networks was the appearance of the deep Residual Network [6] (ResNet). Training up to hundreds or even thousands of layers while achieving compelling performance became possible with it. The key idea behind ResNet is the introduction of the so-called "shortcut connection" that skips one or more layers. Thus, instead of directly fitting a desired underlying mapping, these layers can fit a residual mapping. This leads to better performance and decreases the computational complexity and the process time.

Ronneberger et al. presented a U-shaped network and a training strategy [18] that relies on the strong use of data augmentation for biomedical image segmentation. Their idea was to supplement a usual contracting network by successive layers, where pooling operators are replaced by upsampling operators. Additionally, they concatenated the corresponding feature maps between the expansive and contracting paths to enhance the information flow and consequently increase the resolution of the output.

The backbone of our network is the ResNet-based fully convolutional neural net introduced by Laina et al. [9]. We enhance this network with U-net-like lateral connections to boost its accuracy. The exact modifications are specified in Section 3.

3. Network Architecture

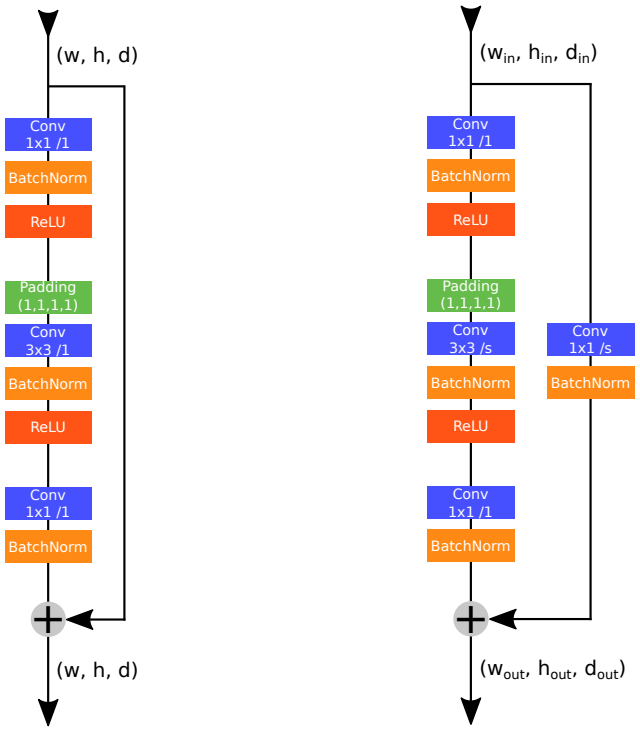


1. ábra. The representations of the different layers used in our architecture. The Up-proj, ResNet-proj and ResNet-skip notations refer to more complex building blocks. See Fig. 2 and 3 for a detailed description. The definition for the interleaving operation can be found in [9].

This section describes the structure of our model. For better understanding, Fig. 1 shows the building blocks we used to assemble the network.

The model outlined in [9] can be divided into two parts: an encoding part based on the ResNet50 architecture, and a decoding part consisting of repeated up-projection blocks (see Fig. 3). Our aim was to boost the reliability of this model without significantly increasing its depth and the number of its variables.

Inspired by U-net-like architectures, instead of deepening the model by increasing the number of layers, we enhance its interconnectedness by introducing side-to-side connections into the network flow. These lateral connections are realized by concatenating the corresponding feature maps between the encoding and decoding parts of the model. After each depth concatenation, we insert an additional 1×1 convolution in order to keep the number of input channels of the up-projection blocks intact. Thanks to these connections the network is able to translate additional information from its previous feature maps into its expansive path. The complete architecture of the proposed network is depicted in Fig. 4.



(a) Residual-skip building block (b) Residual-projection building block

2. ábra

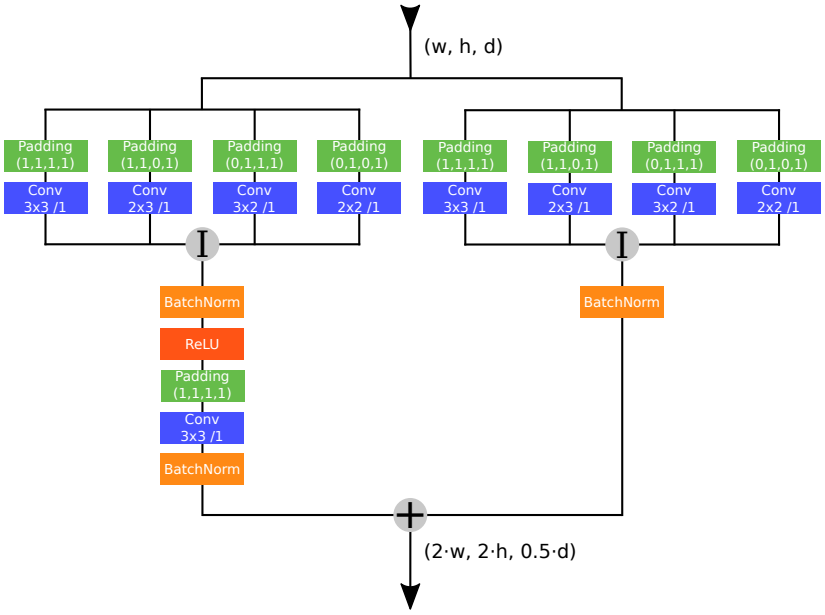
4. Training

4.1. Dataset and Augmentation

We train and evaluate our model on the NYU Depth v2 dataset. This dataset captures 464 indoor scenes with an RGB camera and a Microsoft Kinect. The dataset is split into 215 testing and 249 training scenes. We extracted equally-spaced RGB-D image pairs from the training scenes of the raw dataset and ended up with roughly 48000 pairs as our training set.

Before training, we resize every RGB-D pair, to 352×264 pixels. This way the image sizes are closer to the input shape of our network and the aspect ratio stays roughly the same. Additionally, we use random online augmentation during training. Every RGB-D image pair is:

- scaled by $s \in_R [1, 1.5]$, and the depths are divided by s
- rotated by $r \in_R [-5, 5]$ degree
- randomly cropped down to 320×256 pixels
- the RGB color values are multiplied by a random $c \in_R [0.8, 1.2]^3$



3. ábra. An up-projection block [9]. This block is equivalent to a residual up-projection, but this version is more efficient and leads to reduced training time.

– horizontally flipped with 0.5 probability.

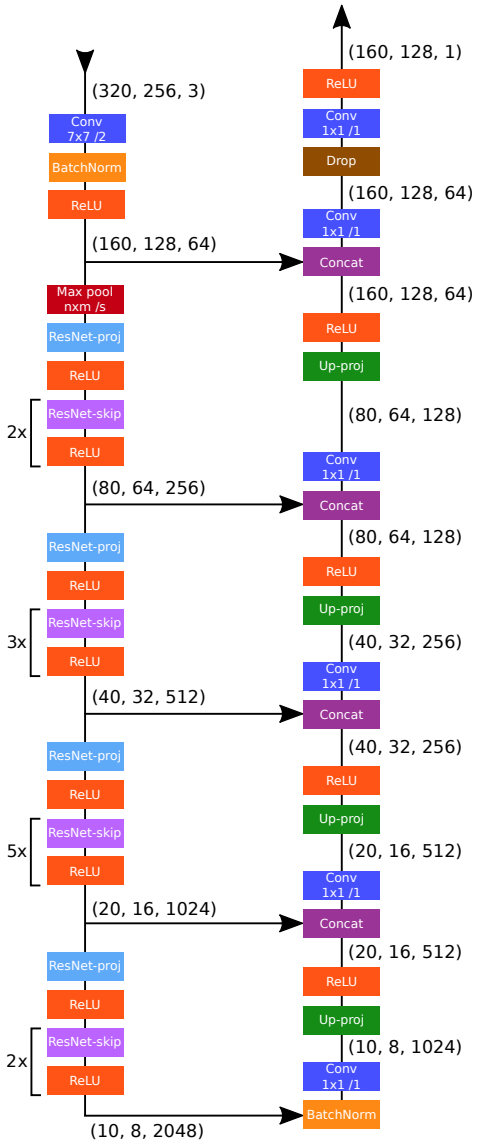
The division by s in the scaling step is necessary to preserve the world-space geometry of the scene. In order to conserve the boundaries between valid and invalid pixels on the depth images, we use Nearest-neighbor interpolation for the scaling and the rotation. After the augmentation, the size of the RGB-D pairs is 320×256 pixels. We further downscale the depth image to 160×128 pixels to match the output size of our network.

It is important to note, that the obtained training dataset contains missing and invalid depth values. We can tackle this problem by masking these invalid depth values during the loss calculation.

4.2. Loss Function

We use reverse Huber (BerHu) loss [14] as our loss function during training instead of the regularly used $\mathcal{L}_2$ loss. This concept was first introduced by Laina et al. [9]. The BerHu loss puts a higher weight on the pixels with higher residuals by applying $\mathcal{L}_2$ on them. Simultaneously, it allows smaller residuals to have a larger effect on the gradients during training due to the use of the $\mathcal{L}_1$ loss.

In order to calculate the BerHu loss $\mathcal{B}(y, \tilde{y})$ for a batch of predictions y and the ground truths $\tilde{y}$, we have to compute $c = \frac{1}{5} \cdot \max_i |y_i - \tilde{y}_i|$, where i indexes



4. ábra. The illustration of our network architecture.

over every pixel of every image in the batch y. Once c is calculated $\mathcal{B}(y, \tilde{y})$ is defined as:

$$\mathcal{B}(y, \tilde{y}) = \begin{cases} |y - \tilde{y}|, & \text{where } |y - \tilde{y}| \leq c. \\ \frac{(y-\tilde{y})^2 + c^2}{2c}, & \text{otherwise.} \end{cases} \tag{1}$$

Therefore, the BerHu loss is equal to the $\mathcal{L}_1$ norm on the pixel i where $|y_i - \tilde{y}_i| \in [-c, c]$ and equal to $\mathcal{L}_2$ outside this interval. The version defined in equation (1) comes from [9]. This form is favorable because it is continuous and differentiable in the switch point c.

4.3. Hyper-Parameter Selection

The first half of the network is responsible for encoding the images. This part is identical to the ResNet-50 network architecture. Thus, we can initialize these layers with the ResNet-50 weights pre-trained on ImageNet [19]. The variables in the second half are initialized by random normal distribution with 0 mean and 0.001 standard deviation.

We train our network for 25-30 epochs, with a batch size of 16. We use a stochastic gradient descent optimizer with 0.9 momentum. The initial learning rate is 10^{-2}. After 10 epochs it is halved, and for the final 5-10 epochs, it is reduced to 10^{-3}.

To prevent overfitting, a dropout layer with a dropout rate set to 0.5 is inserted into the network before the final convolution. Additionally, we set a weight decay of 0.00025 for every layer in our network.

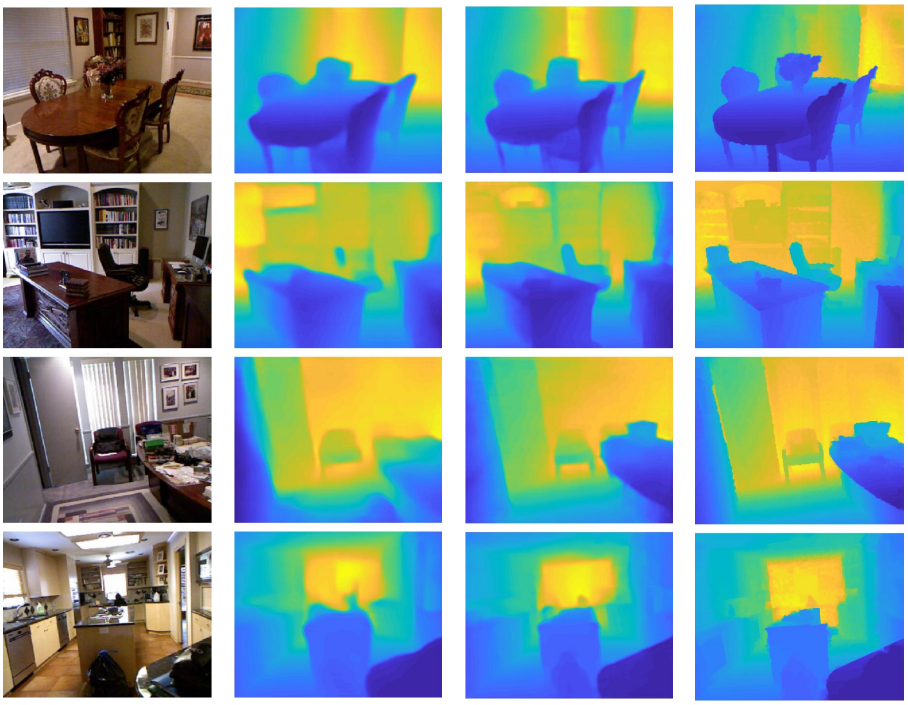
We implemented the network architecture in PyTorch [15], and trained on an NVIDIA GeForce GTX 1080 Ti Graphics Card. An entire training session took approximately 15 hours. Once the model is loaded into the GPU memory, forwarding an arbitrary image through the network takes roughly 0.01 seconds (after resizing the picture to match the input size of the network). This means that our model is fast enough to be utilized in applications that require real-time performance.

1. táblázat. Quantitative comparison with state-of-the-art CNN based methods on the NYU Depth v2 dataset. In the case of RMSE, REL, and Log_{10}, lower is better. For the δ_i accuracies, higher is better.

	RMSE	REL	Log_{10}	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen et al. [3]	0.907	0.215	–	0.611	0.887	0.971
Wang et al. [28]	0.745	0.220	–	0.605	0.890	0.970
Eigen and Fergus [2]	0.641	0.158	–	0.769	0.950	0.988
Laina et al. [9]	0.597	0.137	0.059	0.818	0.955	0.988
Proposed	**0.593**	**0.130**	**0.057**	**0.833**	**0.960**	**0.989**

5. Evaluation

For evaluation, we use the 654 RGB-D image pairs from the labeled test subsets of the NYU Depth v2. We resize the RGB images to 352×264 pixels, and center-



(a) RGB input (b) Laina et al. [9] (c) Proposed model (d) Ground truth

5. ábra

crop them, to match the input size of the model. The output size of the model is 160×128 pixels, while the size of the ground truth depth maps of the test set is 640×480. To compare the results without distortions, we resize the predictions to 582×466 pixels, and center crop the ground truth depth maps to the same size. For a fair quantitative comparison, we apply the same conversions to the output of Laina et al.'s model [9]. The rest of the data in Table 1 is based on the values reported by the authors. We computed the following error metrics on the test set, for quantitative evaluation:

- Root Mean Squared Error: $\sqrt{\frac{1}{n}\sum_{i=0}^{n}(y_i - \tilde{y_i})^2}$
- Relative Absolute Error: $\frac{1}{n}\sum_{i=0}^{n}\frac{|y_i - \tilde{y_i}|}{\tilde{y_i}}$
- Mean Log_{10} Error: $\frac{1}{n}\sum_{i=0}^{n}|\log_{10}(y_i) - \log_{10}(\tilde{y_i})|$
- % of pixels where $max(\frac{y_i}{\tilde{y_i}}, \frac{\tilde{y_i}}{y_i}) = \delta <$ threshold

where i iterates over every pixel of every image of the test set, and n is the number of these pixels. The comparison between our network and other models can be seen in Table 1. For qualitative results, see Fig. 5. Both the qualitative and the quantitative evaluation shows a considerable improvement compared the original architecture as well as the other state-of-the-art approaches. The trained

model is publicly available at `https://github.com/karoly-hars/DE_resnet_unet_hyb`, along with our evaluation code.

6. Conclusion

In this paper, we studied the possibility of boosting the precision of CNNs for depth estimation by inserting lateral connections between the contracting and expansive parts of the networks. We introduced a fully convolutional network which combines the benefits of the residual shortcuts of ResNets and the side-to-side feature map concatenations of U-nets. Our evaluation showed that the proposed network is able to exceed other popular CNN based depth estimation models. In the future we try to design new architectures for the problem addressed at the beginning of the paper. Our aim is to develop an architecture with less storage space requirements and smaller computational complexity in order to use the new network on mobile devices.

Hivatkozások

1. Alexander, E., Guo, Q., Koppal, S., Gortler, S. J., Zickler, T.: Focal Flow: Velocity and Depth from Differential Defocus Through Motion. International Journal of Computer Vision, 1–22 (2017)
2. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. Proceedings of the IEEE International Conference on Computer Vision., 2650–2658 (2015)
3. Eigen, D. and Puhrsch, C. and Fergus, R.: Depth map prediction from a single image using a multi-scale deep network.Advances in neural information processing systems., 2366–2374 (2014)
4. Grossmann, P.: Depth from focus. Pattern recognition letters **5**(1), 63–69 (1987)
5. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from RGB-D images for object detection and segmentation. European Conference on Computer Vision., Springer, 345–360 (2014)
6. He, K., Zhang X., Shaoqing R., Jian S.: Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition 770–778 (2016)
7. Krizhevsky, A., Sutskever I., Hinton G. E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 1097–1105 (2012)
8. Kundu, A., Li, Y., Dellaert, F., Li, F., Rehg, J. M.: Joint semantic segmentation and 3d reconstruction from monocular video. European Conference on Computer Vision. Springer, 703–718 (2014)
9. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. Conference on 3D Vision (3DV), 2016 Fourth International, 239–248 (2016)
10. Li, S. Z.: Markov random field models in computer vision. European conference on computer vision. Springer, 361–370 (1994)
11. Lin, G., Shen, C., Van Den Hengel, A., Reid, I.: Efficient piecewise training of deep structured models for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition., 3194–3203 (2016)

12. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition., 5162–5170 (2015)
13. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition., 3431–3440 (2015)
14. Owen, A. B.: A robust hybrid of lasso and ridge regression. Contemporary Mathematics, **443**(7), 59–72 (2007)
15. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. Machine Learning for Molecules and Materials (2017)
16. Pfister, T., Charles, J., Zissermann, A.: Flowing convnets for human pose estimation in videos. Proceedings of the IEEE International Conference on Computer Vision. 1913–1921 (2015)
17. Ren, X., Bo, L., Fox, D.: Rgb-(d) scene labeling: Features and algorithms. Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. 2759–2766 (2012)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention, Springer 234–241 (2015)
19. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision, **115**(3) 211–252 (2015)
20. Schonberger, J. L., Frahm, J. M.: Structure-from-motion revisited. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4104–4113 (2016)
21. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. European Conference on Computer Vision., Springer 746–760 (2012)
22. Simonyan, K., Zissermann, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint, (2014)
23. Sutton, C., McCallum, A.: An introduction to conditional random fields. Foundations and Trends in Machine Learning, **4**(4), 267–373 (2012)
24. Szirányi, T., Zerubia, J., Czúni, L., Kato, Z.: Image segmentation using Markov random field model in fully parallel cellular network architectures. Real-Time Imaging, **6**(3), 195–211 (2000)
25. Tao, M. W., Srinivasan, P. P., Malik, J., Rusinkiewicz, S., Ramamoorthi, R.: Depth from shading, defocus, and correspondence using light-field angular coherence. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 1940–1948 (2015)
26. Taylor, J., Shotton, J., Sharp, T., Fitzgibbon, A.: The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. Computer Vision and Pattern Recognition (CVPR), IEEE Conference on., 103–110 (2012)
27. Ullman, S.: The interpretation of structure from motion. Proc. R. Soc. Lond. B **203**(1153), 405–426 (1979)
28. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.: Towards unified depth and semantic prediction from a single image. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2800–2809 (2015)
29. Zhang, R., Tsai, P. S., Cryer, J. E., Shah, M.: Shape-from-shading: a survey. IEEE transactions on pattern analysis and machine intelligence **21**(8), 690–706 (1999)