

Edge-preserving Smoothing Regularization for Monocular Depth Estimation

Saqib Nazir

Research Center for Spatial Information (CEOSpaceTech)
University POLITEHNICA of Bucharest (UPB)
 Bucharest, Romania
 saqib.nazir@upb.ro

Daniela Coltuc

Research Center for Spatial Information (CEOSpaceTech)
University POLITEHNICA of Bucharest (UPB)
 Bucharest, Romania
 daniela.coltuc@upb.ro

Abstract—Monocular depth estimation is a fundamental challenge since the foundation of computer vision with many real-world applications. Recently, the introduction of Deep Convolutional Neural Networks (CNN) has brought significant improvements to this particular problem. There are many solutions for scene depth estimation with a focus on obtaining high-quality depth maps from a given RGB image. The insertion of prior information by adding a smoothing regularization has improved the results. However, the smoothing of the surfaces comes together with a certain degradation of the edges. The goal of this paper is to make a comparison between various regularization terms used either in supervised or self-supervised learning methods. In addition to this, we modified the regularization term used currently in self-supervised methods such to work in a supervised manner. The experimental results on NYU-Depth v2 have shown that the regularization based on $L1$ norm of the gradient is the best and the self-supervised modified one outperforms the rest. Finally, rather than relying on common evaluation metrics, we used an additional accuracy measure based on the Steerable Pyramid and Kullback-Leibler divergence (KLD) for edge accuracy of estimated depths that is more sensitive to positional errors of the edges.

Index Terms—Depth estimation, Deep learning, CNN, Steerable Pyramid

I. INTRODUCTION

Single Image Depth Estimation (SIDE) is a challenging task with many real-world applications such as image segmentation, augmented reality, continuous real-time tracking, human-computer interaction, scene recognition, and most recently self-driving cars[1]. In classical methods, various depth cues were used in order to infer depth from single images, such as depth from defocus [2], shape from the shadow, and variation in illumination [3]. However, the estimation of depth from a single image remains a difficult task because in this case, the main cue available is the scene content. For example, in a street image, there are the buildings, cars, footpaths, etc.

Before the advancements of the neural networks in this field, classical appearance-based approaches with Conditional Random Field (CRF) and Markov random field were popular[4, 5]. The recent employment of CNNs has enabled the learning of implicit relation between depth and color pixels. The features learned through CNN are better than conventional handcrafted features. CNN techniques have often been integrated with CRF-based regularization, either as a post-processing step or

via structured deep learning [5]. These methods are more complex, either because of the large number of parameters in a deep network or because of the joint use of a CNN and a CRF [9]. Nevertheless, deep learning has been able to significantly increase the accuracy on standard benchmark datasets, making these methods first in the state of the art. More recently, some studies have significantly improved the performance of estimated depth maps by introducing a smoothing regularization [9, 11, 12, 13, 14, 15]. This loss term works in a complementary manner with other losses and predicts good quality edges in the estimated depth maps.

In this paper, we analyze some existing solutions for edge improvements in order to select the most efficient ones. Although the previous works were able to recover good results for SIDE, we can still observe the distortions in object shape, missing small details, mosaic patterns, and smooth edges. We review a set of smoothing regularization loss functions that helps to generate high-quality smoothing effects by directly learning from data, using CNNs. The heart of the design is the training loss function which includes an edge-preserving regularizer to assist preserve necessary yet potentially vulnerable image structures. We implement and compare several smoothing regularizations used either in the supervised or self-supervised learning methods. We examine the performance of these loss functions on the NYU-Depth V2 dataset [17] and demonstrate that the smoothing regularization based on the $L1$ norm performs better. Finally, we evaluate our results on a better evaluation measure for edge accuracy based on the Steerable Pyramid decomposition.

II. RELATED WORK

In the introductory section, we mentioned some classical and latest techniques based on CNN's for SIDE. In this section, we will address the most recent results in the related studies. Recently several works have used deep learning techniques along with regularization terms in the loss functions to perform the monocular depth estimation. Eigen et al. [9] proposed a supervised multi-scale architecture that predicts the depth of an image on multiple scales. They used $L2$ norm and scale-invariant differences to compare the depths, along with the $L2$ norm of gradients to smooth the result. In [11], Ummerhofer et al. employ point-wise losses for depth, surface normal,

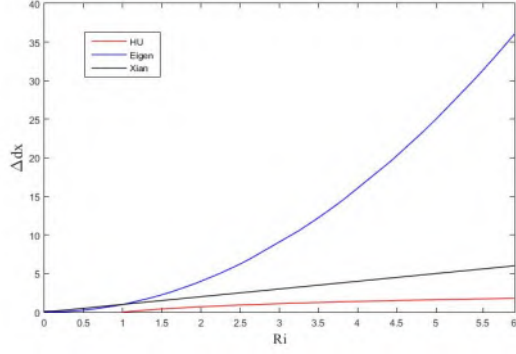


Fig. 1: Smoothing regularization in supervised methods.

and optical flow and smoothing regularization based on the difference of gradients of the estimated and ground-truth maps. The loss is calculated with the $L2$ norm and by considering several scales. Hu et al. [12] propose a supervised method that uses the logarithm of $L1$ norm for both the photometric loss and gradient-based regularization. Recently, the authors of [13] added to a pair-wise ranking loss, a multi-scale gradient matching loss to enforce the smoothing gradients and sharp discontinuities in the supervised method.

In the unsupervised methods where the ground truth is missing, the minimization of the gradient is assisted by an exponential that guarantees the edges are preserved. In [15], the authors use a combination of $L1$ norm and SSIM for the photometric loss function and such gradient for the smoothing regularization. In [10] the authors use in a similar manner the second order gradient. A similar gradient loss to [15], is employed by authors in [14, 16].

Most of the above mentioned studies use U-net architecture based on the Encoder-Decoder network (with skip connections) for SIDE. Pre-trained weights like ResNet50, and DenseNet are most commonly used to initialize the Encoder part of the network. For our experiments, we employed the model outlined in [12], which consists of four parts: an Encoder-Decoder with two extra modules, a multi-scale feature fusion module, and a refinement module.

III. SMOOTHING REGULARIZATION

A. Supervised methods

The smoothing regularization is employed for smoothing the homogeneous areas of the objects in a scene without degrading the edges of a predicted depth. Considering the location of the abrupt edges in the natural images are unknown, they should be identified at the same time as the object is being reconstructed. In order to encourage smoothing within a flat region and discourage smoothing across edges, different solutions have been proposed in the previous studies [9, 11, 12, 13, 14, 15]. They were already evoked in the previous section, now we shall present them in detail.

Hu et al. [12] use a supervised learning method and formulate their training loss as follows:

$$L = L_D + \lambda L_S + \mu L_N \quad (1)$$

where L_D is the depth estimation loss, L_S is the smoothing regularization, and L_N represents the surface normal loss used by [9, 12] to improve the fine details and deal with the small structures of the depth maps. λ, μ are the weighting coefficients. For the depth estimation, they start from the $L1$ norm of the differences between the predicted depth, e^* , and ground truth depth, e , where R_i is defined as:

$$R_i = \text{abs}(e_i^* - e_i) \quad (2)$$

The loss function for depth estimation is:

$$L_D = 1/n \sum_{i=1}^n R_i \quad (3)$$

To overcome the problem of total depth difference that has an equal contribution to the loss between the nearby and distant points in the scene, they use the logarithm of the depth errors:

$$L_D = 1/n \sum_{i=1}^n \log R_i \quad (4)$$

For the smoothing regularization, they consider the log of the gradients in x and y direction:

$$L_S = 1/n \sum_{i=1}^n (\log(\nabla_x(R_i)) + (\log(\nabla_y(R_i))) \quad (5)$$

where $\nabla_x(R_i)$ is the spatial derivative of R_i computed at the i_{th} pixel with respect to x, and $\nabla_y(R_i)$ is the derivative of R_i computed with respect to y. The effect of this smoothing term can be seen in Fig. 1 (Hu).

In [9], the smoothing regularization is done by using the squared $L2$ norm of the gradient:

$$L_S = 1/n \sum_{i=1}^n [(\nabla_x R_i)^2 + (\nabla_y R_i)^2] \quad (6)$$

First-order matching terms $(\nabla_x R_i)^2 + (\nabla_y R_i)^2$ are used in order to obtain predictions that are not only close by values, but also contain similar local structures. In [11] instead of using the gradient of the difference image R_i , the authors provide a solution based on the image normalized gradient:

$$\nabla_h[M](i,j) = \left(\frac{M(i+h,j) - M(i,j)}{|M(i+h,j)| + |M(i,j)|}, \frac{M(i,j+h) - M(i,j)}{|M(i,j+h)| + |M(i,j)|} \right) \quad (7)$$

where $M(i,j)$ are the pixels values in the image. Based on this gradient they define a scale-invariant loss for smoothing regularization:

$$L_S = \sum_{h \in \{1,2,4,8,16\}} \sum_{ij} |\nabla_h[e^*](i,j)| - |\nabla_h[e](i,j)| \quad (8)$$

here h is used to cover gradients at different scales. This loss function encourages the network to match the depth values

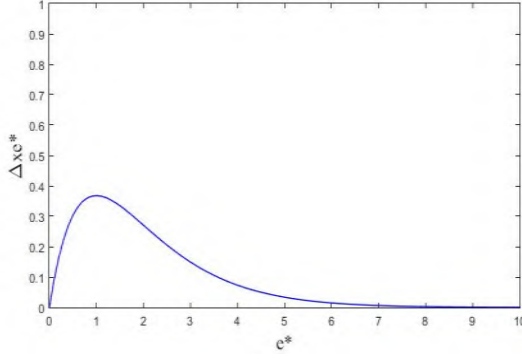


Fig. 2: Smoothing regularization in self-supervised methods.

within a local neighborhood for each pixel. Also, they claim that this loss function increases the smoothness within the homogeneous regions of the predicted depths by highlighting the depth discontinuities and restoring the sharp edges.

A similar multi-scale gradient matching term is used in [13]. They calculate the average sum of the gradients on the x and y-axis on a valid number of pixels.

$$L_S = \frac{1}{n} \sum_s \sum_i (|\nabla_x R_i^s|) + (|\nabla_y R_i^s|) \quad (9)$$

For this smoothing solution, we also consider in our experiments a single scale, similarly as for (8). The effect of this regularization term is shown in Fig. 1 (Xian) is more or less the same as previously described supervised methods, as this loss function penalizes the larger differences, as well.

B. Self-supervised methods

Most of the solutions provided by supervised methods calculate the gradients on the difference between estimated and ground truth depth. Recently, some neural networks based methods have been proposed for depth estimation, which does not require the ground truth depth maps while training. Unlike supervised methods where ground truth depth is available, in self-supervised methods the solution is based on the gradient of the estimated depth e^* [14, 15, 16]:

$$L_S = \frac{1}{N} \sum |\nabla_x e^*| e^{-|\nabla_x I|} + |\nabla_y e^*| e^{-|\nabla_y I|} \quad (10)$$

where I is the RGB input image. The gradient is weighted by an exponential, whose effect is shown in Fig. 2. Since the exponential is sensitive to smaller gradients and insensitive to larger gradients, the prominent edges are preserved. For training a supervised model with such regularization, we replace I with the ground truth e , to take advantage of this a priori information. Last but not least we calculated the results without any regularization term to see the effects on the edges of the estimated depth.

C. Edge Accuracy

The evaluation metrics commonly used in the literature are global measures for the accuracy of the estimated depth maps

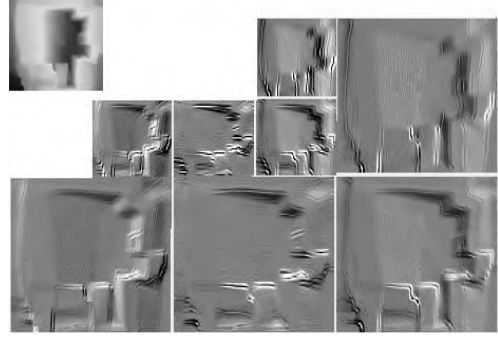


Fig. 3: Steerable pyramid decomposition.

[18]. These metrics have limitations. For instance, most of them are not good enough at identifying the spatial distortion of the object edges [19]. In order to evaluate the accuracy of the edges in the estimated depth maps, we use an additional evaluation measure that is sensitive and more effective to the positional errors of edges.

The Steerable Pyramid: is a linear multi-scale, multi-orientation transform that reveals the edges of the image and allows to measure their quality [20]. The basic functions of the steerable wavelet transform are the K_{th} order directional derivatives, which come in multiple scales and 4 orientations. An example decomposition of the estimated depth map is shown in Fig. 3. This particular steerable pyramid contains 4 orientation subbands, at 2 scales. The reason for selecting the first two scales is that they contain the highest spatial frequency of the image. Knowing that the edge degradation by smoothing is obvious mainly in the domain of high frequencies and considering the size of our test images, we did not proceed to generate lower scales since their contribution to the KLD evaluation would not have been significant.

To evaluate the edges, we decompose both the ground truth and the estimated depth on 2 scales and 4 orientations, resulting 8 sub-bands, then we calculate the KLD on sub-bands. In order to estimate the KLD, we make the histograms of each sub-band and calculate the divergence of corresponding sub-bands as follows [20]:

$$d(h_m||h) = \sum_{i=1}^L h_m(i) \log \frac{h_m(i)}{h(i)} \quad (11)$$

where $h_m(i)$ and $h(i)$ are normalized heights of i_{th} histograms of ground-truth depths and estimated depths, respectively, and L is the number of the bins in the histograms.

Finally, the global KLD between the estimated and ground truth depths is obtained as:

$$D = \log_2(1 + \frac{1}{D_o} \sum_{K=1}^K |d^k(h_m^k||h^k)|) \quad (12)$$

here K is the number of subbands, and D_o is the constant used to control the scale of the distortion measure, in our case $k = 8$ and $D_o = 10$.

IV. EXPERIMENTS

A. Implementation details

For testing the various smoothing regularizations in Section III, we have used the neural network proposed in [12] and the supervised learning. The first half of the proposed network consists of ResNet-50 architecture which is responsible for encoding the images. The encoder part is initialized with pre-trained weights on ImageNet, while the other layers in the network are initialized randomly.

For the training, we use the NYU-Depth V2 dataset [17], which consists of 464 indoor scenes captured with a Microsoft Kinect camera. Since we have taken [12] as the base paper we are considering only indoor images for our experiments. This dataset is most commonly used for the task of SIDE [6, 8, 9, 13, 14, 16]. Following the procedure of [12], we divided the dataset into 249 training and 215 testing scenes. The dataset is down-sampled from (640x480) to (320x240). To fit the size of output depth, the raw depth maps are downsampled to 152x114. For testing, we have used 654 RGB-D samples from the official labeled test subset.

Additionally, we used random online data augmentation during the training. The whole dataset was augmented by color jitter, RGB, and depth image flip and rotations. Like in [12] we trained the model for 20-25 epochs with a batch size of 8. We use Adam optimizer with an initial learning rate of 10^{-3} , and the learning rate is reduced to 10% after every 5 epochs.

We are using the loss function mentioned in Eq. (1) where we replace only the smoothing regularization L_s . We use the modest values for the balancing factors $\lambda, \mu = 0.5$ in our experiments. The network has been implemented using PyTorch and for the steering wavelet transform, we have used an existing implementation in Matlab [20]. The entire training session takes approximately 17 hours on NVIDIA Quadro GV100 GPU with 32 GB memory. Once the trained model is loaded onto the GPU, it takes roughly 0.001 seconds to estimate depth from the given RGB image.

B. Results

Table 1 shows the results obtained for 5 different regularizations and for the case without any regularization. Seven metrics have been used to evaluate quantitatively the results:

- Root mean squared error (RMS)
- Mean relative error (REL)
- Mean log 10 error (log 10)
- Threshold accuracy: $\delta_1, \delta_2, \delta_3$
- Kullback-Leibler divergence (KLD)

It can be seen that the worst results are obtained without the smoothing regularization, which demonstrates once a more the usefulness of introducing such regularization. The best results are obtained for Hu [12] and Xian [13], the later one slightly in advantage except for RMS. Godard [14] overcomes Eigen [9], although it comes from the self-supervised methods.

The results of *KLD* for the edge accuracy are shown on the last column of Table 1. Although supervised methods produced good performance on common evaluation metrics,

their accuracy is not appreciable on edge accuracy except Xian [13], which is by far the best one. Surprisingly, the second best results are provided by the modified solution coming from the self-supervised methods.

The explanation is in using the same L1 norm. It is known from the theory of compressive sensing that this norm is able to find sparse solutions. In smoothing regularization, L1 norm is applied to the gradient of images, which reveals the sparsity.

Figure 5 shows the depth maps estimated by using various smoothing regularizations, and their corresponding RGB and ground truth images. It can be seen that the results without the regularization term suffer from heavy distortion of shapes. Although Hu and Eigen [9, 12] were able to produce images with clear boundaries, they present however several incorrect soft edges, for example, the boundaries of the light lamp on the desk in the case of Hu [12] and objects on the sink in the case of Eigen [9]. The modified regularization term in Godard [14] shows considerable performance by accurately producing edges of the objects and minute structures, such as objects around the sink and light lamp on the desk. Also in the image with the sofa, clear boundaries can be seen. Here, Xian [13] was unable to detect accurate boundaries of this image.

V. CONCLUSION

In this paper, we studied the effects of different smoothing regularization terms on the edges of estimated depth maps generated by CNN's. For experiments, we have used as support a U-net based architecture recently proposed in the literature [12] and dedicated to supervised learning. We tested five different regularization terms based on the image gradient. Four of them have been used in supervised methods whilst the fifth one is the term currently used in the self-supervised approaches. We modified this term such to take as reference the ground truth instead of the RGB image. With our experiments, we show that the methods using regularization terms based on the L1 norm achieve the best accuracy. The quality of edges has been measured by a dedicated metrics consisting in KLD applied on the steering pyramid decomposition of depth maps. The results obtained on images from NYU-Depth V2 dataset show that the regularization used in [13] based on the L1 norm gives the best results. The second best result is given by the regularization coming from the self-supervised area [14]. We believe that this latter solution could be improved by better controlling the sensitivity of loss function to gradients through the exponential, at least in the case of supervised learning. We intend, in a future work, to proceed with our research in this direction and to extend our experiments to the case of outdoor images.

ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860370.

TABLE I: Quantitative comparison with various smoothing regularisations. For RMS, REL, Log 10, and KLD, lower is better. For δ_1 , δ_2 , and δ_3 higher is better.

Method	RMS	REL	Log 10	δ_1	δ_2	δ_3	KLD
Non-Regularization	0.619	0.159	0.064	0.786	0.950	0.986	9.0091e+03
Hu[12]	0.555	0.126	0.054	0.841	0.967	0.991	7.2385e+03
Eigen[9]	0.598	0.143	0.060	0.812	0.956	0.988	7.3216e+03
Ummenhofer[11]	0.582	0.134	0.058	0.822	0.963	0.991	7.2817e+03
Xian[13]	0.559	0.125	0.054	0.844	0.968	0.991	7.1009e+03
Godard[14] (Modified)	0.575	0.142	0.057	0.825	0.961	0.988	7.1929e+03

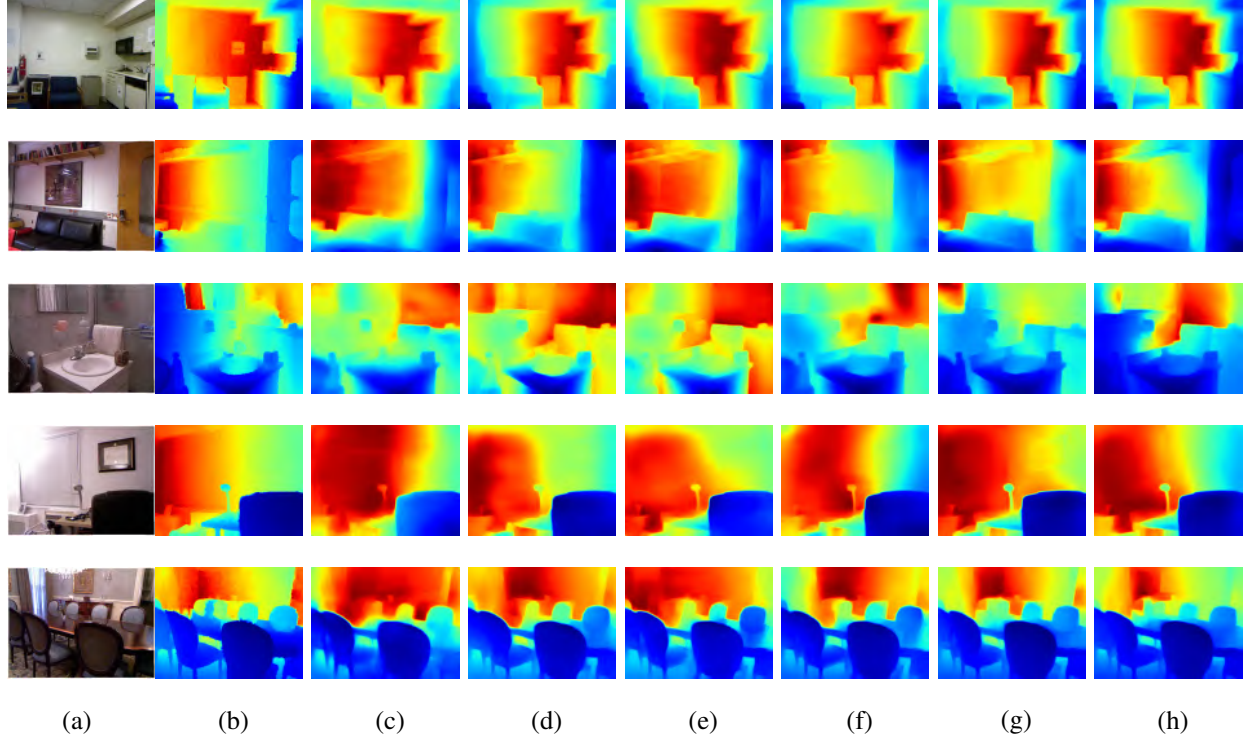


Fig. 4: Estimated depth maps for 6 scenes and 5 different smoothing regularizations. From the first to the last row; (a) input RGB images, (b) ground truth depth map, (c) results without regularization term, (d) Hu [12], (e) Eigen [9], (f) Ummenhofer [11], (g) Xian [13], and (h) Godard [14] (modified) loss function.

REFERENCES

- [1] Khan, Faisal, Saqib Salahuddin, and Hossein Javidnia. "Deep learning-based monocular depth estimation methods—A state-of-the-art review." *Sensors* 20.8 (2020): 2272.
- [2] Maximov, Maxim, Kevin Galim, and Laura Leal-Taixé. "Focus on defocus: bridging the synthetic to real domain gap for depth estimation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
- [3] Zhao, ChaoQiang, et al. "Monocular depth estimation based on deep learning: An overview." *Science China Technological Sciences* (2020): 1-16.
- [4] Liu, Jun, et al. "A contextual conditional random field network for monocular depth estimation." *Image and Vision Computing* 98 (2020): 103922.
- [5] Lu, Yawen, Michel Sarkis, and Guoyu Lu. "Multi-Task Learning for Single Image Depth Estimation and Segmentation Based on Unsupervised Network." 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020.
- [6] 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)*, 2014.
- [7] B. Li, C. Shen, Y. Dai, A. V. den Hengel, and M. He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 1119–1127, 2015.
- [8] Cao, Yuanzhouhan, Zifeng Wu, and Chunhua Shen. "Estimating depth from monocular images as classification using deep fully convolutional residual networks." *IEEE Transactions on Circuits and Systems for Video Technol-*

ogy 28.11 (2017): 3174-3182.

- [9] Eigen, David, and Rob Fergus. "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture." *Proceedings of the IEEE international conference on computer vision*. 2015.
- [10] Wang, Chaoyang, et al. "Learning depth from monocular videos using direct methods." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
- [11] Ummenhofer, Benjamin, et al. "Demon: Depth and motion network for learning monocular stereo." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
- [12] Hu, Junjie, et al. "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries." *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019.
- [13] Xian, Ke, et al. "Structure-guided ranking loss for single image depth prediction." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
- [14] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
- [15] Godard, Clément, Oisín Mac Aodha, and Gabriel J. Brostow. "Unsupervised monocular depth estimation with left-right consistency." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
- [16] Gur, Shir, and Lior Wolf. "Single image depth estimation trained via depth from defocus cues." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.
- [17] Silberman, Nathan, et al. "Indoor segmentation and support inference from rgb-d images." *European conference on computer vision*. Springer, Berlin, Heidelberg, 2012.
- [18] Hambarde, Praful, and Subrahmanyam Murala. "S2dnet: Depth estimation from single image and sparse samples." *IEEE Transactions on Computational Imaging* 6 (2020): 806-817.
- [19] Zhu, Shengjie, Garrick Brazil, and Xiaoming Liu. "The edge of depth: Explicit constraints between segmentation and depth." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
- [20] Wang, Zhou, and Eero P. Simoncelli. "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model." *Human vision and electronic imaging X*. Vol. 5666. International Society for Optics and Photonics, 2005.
- [21] Ramamonjisoa, Michaël, Yuming Du, and Vincent Lepetit. "Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.