Laplacian Pyramid-based Depth Residuals for Self-Supervised Monocular Depth Estimation

E In Son Myung-woo Woo Jihoon Hwang
Department of ECE, Seoul National Unviersity, Seoul, Korea
{pingpang, wmw2000, hoons21}@snu.ac.kr

Abstract

With advances in deep convolutional neural networks (DCNNs), monocular depth estimation has shown very promising results. However, most existing methods handle depth estimation as a supervised regression problem, which suffers from acquiring per-pixel ground-truth depth data at scale. To overcome this limitation, recent works approached in a self-supervised manner, by addressing the monocular depth prediction task as a reconstruction problem. In this paper, we propose three new ideas to enhance self-supervised monocular depth estimation: 1) laplacian pyramid, 2) receptive field block and 3) brightness consistency loss. Laplacian pyramid incorporated in the decoder architecture can successfully emphasize the difference across the scale spaces, which can precisely estimates the depth boundary as well as the global layout. Receptive field block on encoder architecture gives a great help to incorporate more discriminative feature representation and improve flow of gradients. Brightness consistency loss is designed to relax the luminance difference between frames. Empirical evaluation on the KITTI dataset demonstrates the effectiveness of our approach.

1. Introduction

Depth estimation from 2D images has been studied in computer vision for long time and nowadays applied to robotics, autonomous driving cars, 3D reconstructions and scene understanding. Those approaches usually relied on multiple instances of the same scene such as stereo image pairs, multiple frames from moving camera, or static captures under different lighting conditions. As depth estimation from multiple observations have enjoyed great success, it naturally lead to depth estimation with a monocular image since it demands less cost and constraint.

Depth estimation from a single image have been limited by traditional approaches because it is an inherently ill-posed problem. However, deep learning has achieved remarkable growth by learning not only image features but also peripheral information such as camera pose, optical flow, surface normal, segmentation etc., to approach the level of performance obtained by binocular images. Most existing methods treat monocular depth estimation as a supervised regression problem and as a result, require vast quantities of ground-truth data for training, which is very costly. For instance, in the scenario of depth estimation for autonomous driving, it implies driving a car equipped with a laser LiDAR scanner for hours under diverse lightning and weather conditions. Self-Supervised methods replace the used of explicit depth data during training with easier-to-obtain synchronized stereo pairs [7] or monocular video [30]. By hallucinating the depth for a given image and projecting it into nearby views, the model is trained by minimizing the image reconstruction error. In other words, self-supervised methods treat depth estimation as an image reconstruction problem.

Figure 1: **Depth from monocular image.** Our self-supervised model produces sharp, high quality depth maps.

In this paper, we propose three architectural and loss innovations for self-supervised monocular depth estimation. 1) laplacian pyramid 2) receptive field block and 3) brightness consistency parameters. Our proposed laplacian pyramid precisely interprets the relation between the encoded features and the final output for monocular depth estima-

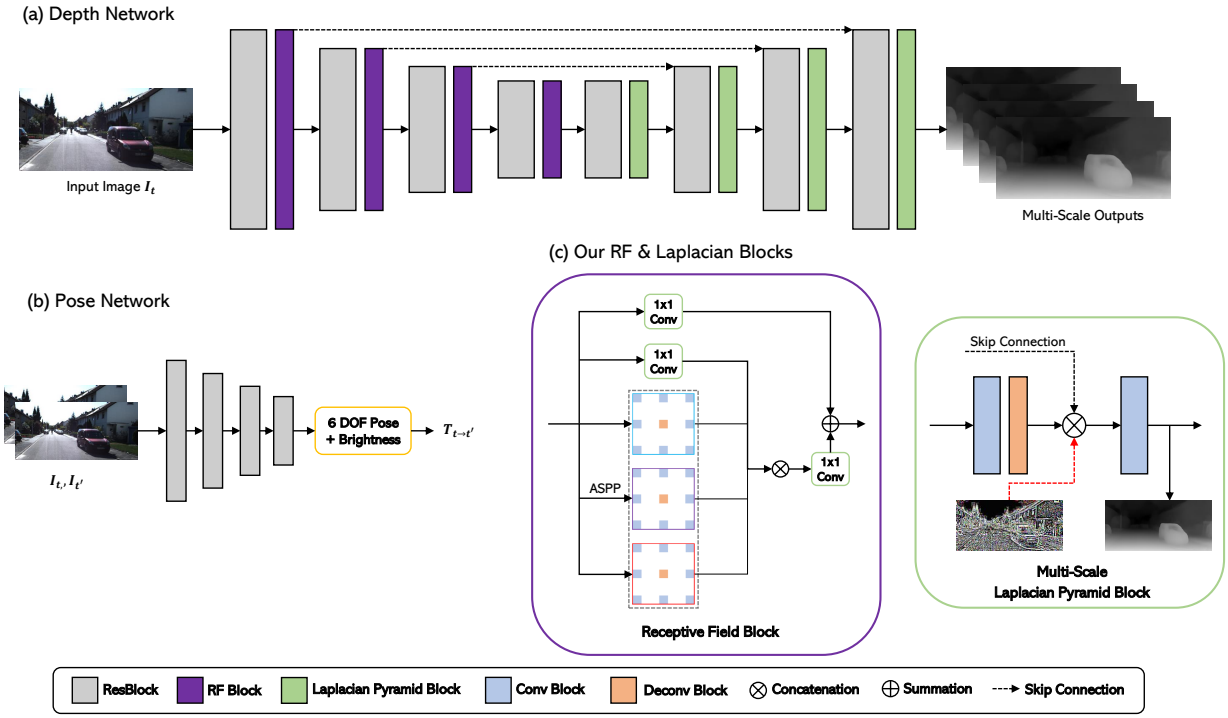

Figure 2: **Overview of the proposed network architecture. (a) Depth Network:** The input monocular image is encoded using ResNet with RF blocks. The encoded features are fed into stacked convolution blocks to generate sub-band depth residuals at each pyramid level. **(b) Pose Network:** Pose and brightness difference between a pair of frames is predicted with a seperate pose network. **(c) RF & Laplacian Blocks:** The RF block efficiently capture encoded features with large receptive field. The ASPP module consists of 3 dilated convolutional layers with kernel size and dilated rate 3, 5, 7 respectively. Laplacian pyramid block gives a guide to decoding process with residuals of the input color image with the depth residuals at each pyramid level.

tion. Laplacian has been used in various fields of scene understanding because of its ability to preserve the local information of the given data [13]. We exploit laplacian pyramid-based decoder architecture, which is highly relevant to object boundaries, to precisely interpret the relation between the encoded features and the final output for monocular depth estimation. Specifically, the encoded features are fed into stacked convolution blocks to generate sub-band depth residuals at each pyramid level. Depth network consists of basic encoder-decoder structure. Good encoding is necessary to produce good output. Of course, ResNet [10] has been verified as a good feature extractor in many fields. ResNet [10] consists of a single kernel structure of 3x3, which is somewhat unfortunate in the collection of multiple scale information through the receive field extension. To compensate for this, we applied a module called RFB [12] between ResNet [10] Feature steps. RFB [12] modules are applied with kernel and dilation of various sizes, expanding the Receive Field while minimizing the increase in parameters. If stereo images are processed as input, the image-pair is synchronized in time. There is a temporal difference between frames when the video(sequence of image) becomes input. When we compute reprojection error, we assume brightness consistency. However, due to the aforementioned temporal difference, brightness changes occur frequently in real environments.These differences result in unnecessary loss. A brightness consistency parameter [26] is a parameter learned within a network that compensates for the brightness difference between adjacent frames. Before computing the final loss, the brightness of the reconstructed image is corrected by the parameters to minimize the occurrence of unnecessary loss. Examples of depth estimation by the proposed method are shown in Fig. 1 The whole network is trained in an end-to-end manner without stage-wise training or iterative refinement. Experiments on the KITTI dataset [6].

The remainder of this paper is organized as follows. A comparative review of related works is presented in Section 2. The proposed network is explained in detail in Section 3. In Section 4 experimental results are demonstrated on the

KITTI dataset. The conclusion follow in Section 5.

2. Related Work

Monocular depth estimation is an inherently ill- posed problem. Early trial for estimating depth from single image used had-made feature and probability graphical model. These days, most are trying using modern deep learning.

2.1. Supervised Depth Estimation

According to same input image, there are multiple possible depths. To address this problem, a learning-based method has been used and has had many successes. There have been many attempts to apply end-to-end supervised learning to depth [3], [5], [11]. The accuracy was increased by using various ideas as well as the structure of the network. For example, attempts are made to turn depth estimation into an ordered regression problem. [5].

However, fully supervised learning has fundamental limitations. That is, it requires an elaborately crafted ground truth. Because, obtaining good quality depth information is costly process. As a result, many attempts have been made for partial or weak level supervised learning. *e.g*. Using known object size [25], learning ordinal depths [31], or using synthetic depth [16], all while these methods still require additional information, which are a little cheaper than fully-supervised but still costly.

2.2. Self-Supervised Depth Estimation

Self-supervised depth estimation does not require ground truth depth. One way to train depth estimation without ground truth depth is to learn from image reconstruction. These models use stereo image pairs as input, or continuous monocular image frame.In the training , the other images are reconstructed using only one of the sequence images or the paired images. At this time, errors are reduced by training the disparity.

Self-Supervised Stereo Training

One method in self-supervised learning is to use synchronous stereo-pairs. The synchronized stereo-pairs itself contain depth information. The relationship between disparity and depth is linear. Therefore, disparity is itself information directly about depth. This is how a binocular camera measures depth, the same principle as a human detects depth. In the training, the model learns disparity by the reconstructing image loss.

There have been many attempts, and recently, show superior result by adding a left-right consistency term [7].The method based on stereo is expanding in several directions. There are attempts to use GAN [18] or another consistency [20]. There are also attempts for real-time inference [19]. Although there are many improvement in terms of data collection through self-supervised stereo training, the need for a stereo image for monocular depth estimation is still inefficient.

Self-Supervised Monocular Training

Self-supervised monocular training uses monocular sequence image as input information. That's just using the video taken with a monocular camera. This video contains temporal information. For depth estimation through video, information about the camera pose is also required. The good news is that pose estimation is only needed in the training phase.

Monocular depth estimation has several assumptions. The objects in the input are stationary, and the same object has the same shape. However, in the real environment, changes in light also occur and there are moving objects. Early models of self-supervised monocular training suffered from these challenge [30]. The time difference of monocular video created more challenges to solve than stereo, which resulted in lower performance. Recently, the performance gap between self-supervised monocular and stereo is narrowing. many techniques have been used to mask non-rigid scenes to improve performance [23]. Various methods have been tried, such as strengthening the relationship of edges [27], using a depth normalization layer [24], or creating a mask using pre-trained instance segmentation [1].

2.3. Laplacian Pyramid

Depth decoder utilizes a deconvolution [17] network. This is a learnable upsampling kernel. Although it is much more accurate than upsampling by simple interpolation, it is still difficult to expand compressed information again. Therefore, to supplement this part, it is very common to use skip-connection.

From a similar point of view, laplacian pyramid was used to convey more information about the boundary of objects. Laplacian has been used in various fields of scene understanding because of its ability to preserve the local information of the given data [13]. The laplacian filter is known to extract high-frequency signals well in the field of traditional computer vision. Considering the high frequency signal in terms of the image, the pixel value change is a big part, and this is the boundary of the object. The laplacian pyramid information processed in the encoder stage is progressively delivered to the decoder as a skip-connection.

2.4. Receptive Field Block

ResNet, which is often used as an encoder, uses only a 3x3 kernel. This is not a good method from the point of view of the Receptive Field that collects information of various scales.

GoogleNet [22], known as Inception Block, acquires

multi-scale information using various kernel sizes. However, the use of a large size kernel leads to an increase in the number of parameters. This leads to an increase in the amount of computation. For ASPP [9], dilated convolution introduces the concept of dilation to secure a wide receptive field without a large increase in parameters. However, ASPP causes a lot of information to be lost due to uniform kernel application.

In order to minimize the information lost, RFB [12] uses various size of kernels according to the dilation to minimize the information lost while expanding the RF (Receptive Field).

3. Method

We will focus on the main idea we present: 1) Laplacian pyramid based depth estimation 2) Receptive Field Block 3) Brightness Consistency parameters. Monodepth2 [8] is used as the baseline.

3.1. Baseline:Monodepth2

Monodepth2 [8] is trained by estimating the target image as a different view point. The core concept of the monocular depth estimation network is the self-supervised training scheme which simultaneously learns depth with DepthNet and motion with PoseNet using video sequences. The loss in this network is designed to minimize the photometric reprojection error L_p. The relative pose of the target image I_t for the source image $I_{t'}$ is expressed in $T_{t'\to t}$.

$$\begin{aligned} L_p &= \mathbf{min_{t'}}\, pe(\mathbf{I_t}, \mathbf{I_{t'\to t}}) \\ I_{t'\to t} &= I_{t'} < proj(D_t, T_{t'\to t),K} > \end{aligned} \tag{1}$$

pe is a photometric reconstruction error. $proj()$ are estimated 2D coordinate by Depth D_t in source image $I_{t'}$. $<>$ is the sampling operator. K is pre-computed camera intrinsics K. Both SSIM and L1 distance were used to calculate pe.at each pixel, instead of averaging the photometric error over all source images, we simply use the minimum(see Fig. 2).

$$\begin{aligned} pe(I_{a'}, I_b) &= \frac{\alpha}{2}\left(1 - SSIM\left(I_a, I_b\right)\right) + (1-\alpha)\left\|I_a - I_b\right\|_1 \\ L_s &= \left|\partial_x d_t^*\right| e^{-\left|\partial_x I_t\right|} + \left|\partial_y d_t^*\right| e^{-\left|\partial_y I_t\right|} \end{aligned} \tag{2}$$

where =0.85. As in [7] we use edge-aware smoothness where $d_t^* = d_t/\bar{d}_t$ is the mean-normalized inverse depth from [24] to discourage shrinking of the estimated depth. For monocular training, we use the two frames temporally adjacent to I_t as source frame, $i.e.$ $I_{t'} = \{I_{t-1}, I_{t+1}\}$.

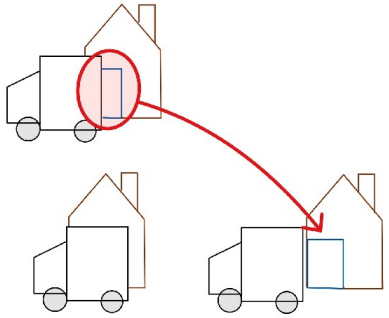

Figure 3: **The reprojection with adjacent frame.**

In addition, Auto-masking Stationary Pixel and Multi-scale Estimation were added to improve performance. Mask objects with motion in the video to prevent inaccurate loss calculations. [ ] is the Iverson bracket. When calculating the final loss, multiply by pe to make a mask.

$$\mu = [\mathbf{min_{t'}}\, pe(I_t, I_{t'\to t}) < \mathbf{min_{t'}}\, pe(I_t, I_{t'})] \tag{3}$$

Final Training Loss

The final loss is computed by applying per-pixel smoothness and Auto-mask to the photometric loss and by applying the average to the pixel, scale, and batch.

$$L = \mu L_p + L_s \tag{4}$$

3.2. Laplacian Pyramid-Based Depth Estimation

To extract information about the boundary of the image, we apply a Laplacian filter to the initial input image.Features extracted from the Laplacian filter are passed through skip-connections to give guidelines to the decoder. To give guidelines to all stages of decoder, downsample to the decoder output size. In summary, the output of the decoder is concatenated with the downsampled Laplacian feature and an feature is passed through skip-connection in the encoder. This combined feature is passed on to the next stage of decoder(see Fig.2)

3.3. Receptive Field Block

ResNet itself is a very good Encoder, but RFB [12] is used to encode multi-scale information by a wide reception field during the Encoding phase. RFB has been applied to encode Feature in all phases (5 stage) of ResNet(see Fig. 2).The RFB's internal structure branches into one 1x1, 3x3, 5x5, 7x7 kernels and one short cut. Each branch then broadens the receiveive field through a kernel size dilation $.e.g.$ 3x3 kernel; 3 dilation, 5x5 kernel; 5 dilation. Finally, concatenate it and make it the same as the original channel via 1x1 conv, sending the ResNet at the next stage. Apply Batch Norm between convs to minimize optimization problems.

Method	Train	lower is better				higer is better		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhou [30]	M	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Yang [28]	M	0.182	1.481	6.501	0.267	0.725	0.906	0.963
Mahjourian [15]	M	0.163	1.240	6.220	0.250	0.762	0.916	0.968
GeoNet [29]	M	0.149	1.060	5.567	0.226	0.796	0.935	0.975
DDVO [24]	M	0.151	1.257	5.583	0.228	0.810	0.936	0.974
DF-Net [32]	M	0.150	1.124	5.507	0.223	0.806	0.933	0.973
LEGO [27]	M	0.162	1.352	6.276	0.252	-	-	-
Ranjan [21]	M	0.148	1.149	5.464	0.226	0.815	0.935	0.973
EPC++ [14]	M	0.141	1.029	5.350	0.216	0.816	0.941	0.976
Struct2depth [1]	M	0.141	1.026	5.291	0.210	0.845	0.948	0.977
Monodepth2 [8]	M	0.115	0.924	4.852	0.193	0.876	0.958	0.981
Monodepth2 + Laplacian	M	0.113	**0.876**	**4.796**	**0.191**	**0.881**	**0.959**	0.981
Monodepth2 + RF	M	**0.112**	0.904	4.851	0.195	0.877	0.958	0.979
Monodepth2 + Brightness	M	0.117	0.890	4.888	0.194	0.870	**0.959**	**0.982**
Ours	M	**0.111**	**0.870**	**4.796**	**0.191**	**0.881**	**0.959**	**0.982**

Table 1: **Quantitative results.** Comparison of our method to existing methods and baseline method on KITTI 2015 [6] using the Eigen split. Best results in each category are in **bold**, with second best results underlined. (**M** - Self-supervised mono supervision)

Figure 4: I_t **Examples of brightness affine transform.** From top to bottom: input image, calibrated input image, source image by brightness calibration.

3.4. Brightness Consistency Parameters

The photometric reprojection error is based on the brightness constancy assumption. However, it can be violated due to illumination changes and auto-exposure of the camera to which both L1 and SSIM are not invariant. Therefore, we propose to explicitly model the camera exposure change wight predictive brightness transformation parameters. The change of the image intensity due to the adjustment of camera exposure can be modeled as an affine transformation with two parameters a, b.

$$I^{a,b} = aI + b \tag{5}$$

Despite its simplicity, this formulation has been shown to be effective in [4]which builds upon the brightness constancy assumption as well. we propose predicting the transformation parameters a, b which align the brightness condition of source images $I_{t'}$ with input image I_t. We reformulate photometric error equation with brightness transformation parameters.

$$L_p = \mathbf{m}in_t pe(I_t, I^{a,b}_{t'\to t}) \tag{6}$$

with

$$I^{a,b}_{t'\to t} = aI_{t'\to t} + b \tag{7}$$

where a and b are the transformation parameters aligning the illumination of I_t to $I_{t'}$. Both parameters can be trained in a self-supervised way without any supervisional signal(see Fig. 3).

4. Experiments

Here, we validate that (1) our Laplacian Pyramid method (2) our RF block method (3) our brightness consistency loss method compared to baseline method. We evaluate our models, named Monodepth3, on the KITTI 2015 [6] stereo dataset, to allow comparison with baseline method.

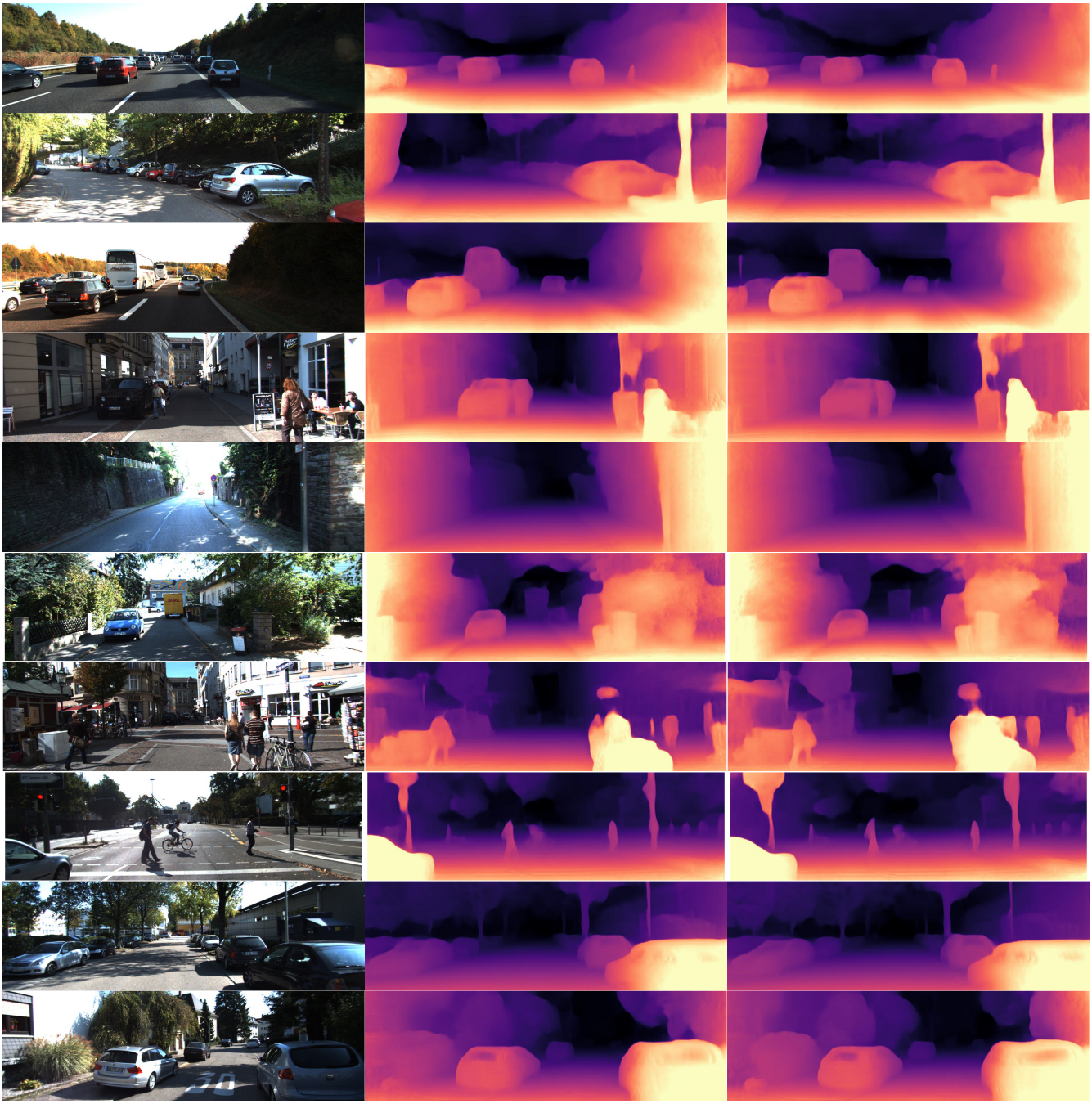

Figure 5: **Qualtative results on the KITTI Eigen splits.** Left: input image; Middle: Monodepth2; Right: Ours

4.1. KITTI Eigen Split

We use the data split of Eigen et al. [2] 39,810 for training and 4,424 for validation. We use the same instrinsics for all images, setting the principal point of the camera to the image center and the focal length to the average of all the focal lengths in KITTI. During evluation, we cap depth to 80m per standard practice[7]. For our monocular models, we report results using the per-image median ground truth scaling introduced by [30].

We compare the results of several variants of our model, trained with monocular video only(M). The upper part of Table 1. shows the comparison with existing models and baseline model which trained with monocular setting. The results demonstrate that the proposed depth estimation network outperforms Monodepth2 on majority metrics. Quali-

tative results can be seen in Fig. 4.

4.2. Benefits of Laplacian Pyramid

Laplacian pyramids extract high frequency components of input data. By recovering depth residuals from encoded features in different levels of laplacian pyramid, the proposed method successfully restores local details $e.g$ depth boundary as well as global layout.

4.3. Benefits of RF

Receptice Field Block uses various size of kernels to minimize the information lost. By recovering depth residuals in different size of kernels on encoder, the proposed method successfully minimize the information lost.

4.4. Benefits of Brightness Consistency Parameters

The full Eigen KITTI split data does not contain large illumination change at adjacent frames. So, our additional Brightness Consistency parameters affect very little improve at KITTI Dataset. However, our method slightly outperform the baseline method.

4.5. Implementation details

batch size = 12, learning rate = 0.0001, number of epochs = 20, number of resnet layers = 18, image = $192x640$, disparity smoothness = 0.001, min depth =$0.1m$, max depth= $100m$, baseline method = $Automasking$, $Multi-scaleequalsizeEstimation$

5. Conclusion

Here, we validate that (1) our Laplacian Pyramid (2) our Receptive filed block (3) our brightness consistency loss compared to baseline reprojection loss. We evaluate our models, named Monodepth3, on the KITTI 2015 stereo dataset, to allow comparison with baseline method.

References

[1] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8001–8008, 2019.

[2] 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*, pages 2650–2658, 2015.

[3] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. *arXiv preprint arXiv:1406.2283*, 2014.

[4] J. Engel, J. Stückler, and D. Cremers. Large-scale direct slam with stereo cameras. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1935–1942. IEEE, 2015.

[5] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018.

[6] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.

[7] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[8] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.

[12] S. Liu, D. Huang, et al. Receptive field block net for accurate and fast object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 385–400, 2018.

[13] W. Liu, X. Ma, Y. Zhou, D. Tao, and J. Cheng. p-laplacian regularization for scene recognition. *IEEE transactions on cybernetics*, 49(8):2927–2940, 2018.

[14] C. Luo, Z. Yang, P. Wang, Y. Wang, W. Xu, R. Nevatia, and A. Yuille. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2624–2641, 2019.

[15] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5667–5675, 2018.

[16] N. Mayer, E. Ilg, P. Fischer, C. Hazirbas, D. Cremers, A. Dosovitskiy, and T. Brox. What makes good synthetic training data for learning disparity and optical flow estimation? *International Journal of Computer Vision*, 126(9):942–960, 2018.

[17] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[18] A. Pilzer, D. Xu, M. Puscas, E. Ricci, and N. Sebe. Unsupervised adversarial depth estimation using cycled generative networks. In *2018 International Conference on 3D Vision (3DV)*, pages 587–595. IEEE, 2018.

[19] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia. Towards real-time unsupervised monocular depth estimation on cpu. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5848–5854. IEEE, 2018.

[20] M. Poggi, F. Tosi, and S. Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *2018 International conference on 3d vision (3DV)*, pages 324–333. IEEE, 2018.

[21] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M. J. Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12240–12249, 2019.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[23] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017.

[24] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2022–2030, 2018.

[25] Y. Wu, S. Ying, and L. Zheng. Size-to-depth: a new perspective for single image depth estimation. *arXiv preprint arXiv:1801.04461*, 2018.

[26] N. Yang, L. v. Stumberg, R. Wang, and D. Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1281–1292, 2020.

[27] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia. Lego: Learning edge with geometry all at once by watching videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 225–234, 2018.

[28] Z. Yang, P. Wang, W. Xu, L. Zhao, and R. Nevatia. Unsupervised learning of geometry with edge-aware depth-normal consistency. *arXiv preprint arXiv:1711.03665*, 2017.

[29] Z. Yin and J. Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018.

[30] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017.

[31] D. Zoran, P. Isola, D. Krishnan, and W. T. Freeman. Learning ordinal relationships for mid-level vision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 388–396, 2015.

[32] Y. Zou, Z. Luo, and J.-B. Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 36–53, 2018.