P2D: a self-supervised method for depth estimation from polarimetry

Marc Blanchon[1†], Désiré Sidibé[2], Olivier Morel[1], Ralph Seulin[1], Daniel Braun[1] and Fabrice Meriaudeau[1]
[1]ERL VIBOT CNRS 6000, ImViA, Université Bourgogne Franche-Comté, 71200, Le Creusot, France
[2]IBISC, Univ Evry, Université Paris-Saclay, 91025, Evry, France
[†]Email: fr.marc.blanchon@gmail.com

***Abstract*—Monocular depth estimation is a recurring subject in the field of computer vision. Its ability to describe scenes via a depth map while reducing the constraints related to the formulation of perspective geometry tends to favor its use. However, despite the constant improvement of algorithms, most methods exploit only colorimetric information. Consequently, robustness to events to which the modality is not sensitive to, like specularity or transparency, is neglected. In response to this phenomenon, we propose using polarimetry as an input for a self-supervised monodepth network. Therefore, we propose exploiting polarization cues to encourage accurate reconstruction of scenes. Furthermore, we include a term of polarimetric regularization to state-of-the-art method to take specific advantage of the data. Our method is evaluated both qualitatively and quantitatively demonstrating that the contribution of this new information as well as an enhanced loss function improves depth estimation results, especially for specular areas.**

I. Introduction

Depth estimation represents an important subject in the field of computer vision, and numerous methods were developed involving geometry and optimization processes. At present, algorithms tend to be artificially lightened by the use of deep learning [1]–[3]. Moreover, a trend towards the relaxation of some constraints can be observed typically with the mono-depth approaches [4]–[6]. One of the common points of all these modern algorithms is they predominantly use a single dataset: KITTI [7], [8]. This dataset is composed of a massive number of RGB images which greatly links the sensory system to the applications. As a result, the modality limitations are exported to the applications and some problematic aspects such as specular surfaces are neglected. By following the logic of depth estimation via a monocular camera, it is possible to provide a new modality able to address the weaknesses of RGB data. In this perspective, polarimetry seems to represent an excellent candidate for depth estimation in urban areas. Indeed, by definition, polarimetric data are particularly able to cope with light phenomena. Moreover, the applications of polarimetry were previously highly specialized because of the cost of the sensors and the specificity of the very constraining acquisition. Nowadays, modern devices tend to both reduce their cost and facilitate dynamic acquisition via focal plane division sensors which induce increasingly popular uses [9]–[12]. In addition, it has been established a deep learning network requires a large amount of diverse data to learn relevant representations. As follows, the scarcity of polarimetric datasets tends to encourage the use of monocular depth estimation methods via self-supervised deep learning.

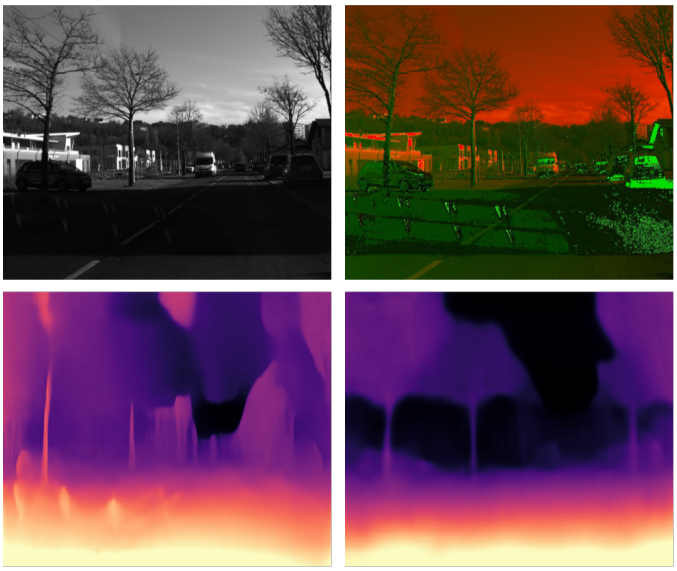


Fig. 1. Illustration of the proposed method compared to a state-of-the-art method [13]. On the left are the input grayscale image and the corresponding disparity obtained by [13]. On the right are the polarimetric input and the result of the proposed P2D.

In this paper, we present P2D (Polarimetry to Depth), a self-supervised monocular depth estimation network taking advantage of polarization cues. As shown in Figure 1, we exploit the polarimetric information to make the estimation sensitive to specularity as illustrated by the plane reconstruction on the car windows. In addition, since the algorithms are already efficient for diffuse areas, we propose not to penalize these estimates but to improve their weaknesses. Since polarimetric data is scarce, we exploit the faculties of self-supervision to train on urban dataset acquired with polarization sensor for the occasion. The particular issue we address is very challenging as it confronts the problems of specularity in urban scenes. We propose an evaluation of a recent deep learning based technique and of our method on a benchmark allowing a common and fair comparison.

II. Related Works

This section will discuss some previous works related to the two main components of this contribution.

Polarimetry and geometry. Polarimetry is directly related to scene geometry. As light interacts with materials, this unique modality acquires the relationships between the two. This property is unrestrictive since this type of data does not uniquely define specular areas but also diffuse areas. Therefore, the use of such modality could be profitable since it provides another complementary information to RGB. Moreover, when a reflection can be clearly defined by polarization, the RGB values tend to saturate or mislead algorithms.

Many algorithms take advantage of polarimetric information such as: Structure from Polarization [10], [14], [15], classification of urban scenes [16] and pose estimation [11], [17]. Also, while using existing methods adapted to this modality, some contributions propose improved versions of scene geometry estimation. In [9], an efficient stereovision method is proposed, reconstructing specular areas by minimizing a cost function inspired by [18]. Other approaches induce concepts of multiple view geometry and tend to require a more constraining acquisition system as in [10], [19]. Ultimately, some methods propose alternatives requiring either an increased knowledge of the scene via the refractive index [20] or an improved acquisition system [21], [22]. Despite an increasing diversity of reconstruction methods with polarization cues, the approaches are still ruled by Fresnel equations linking polarization, materials and normals. In addition, there is a tendency to use polarimetry either as an addition to other modalities or in complex multi-camera/image systems.

It is therefore clear that polarimetry remains a legitimate contender when it comes to taking advantage of scene geometry. Especially since most real-world acquisitions are uncontrolled and therefore scenes are prone to contain specular areas. It is also noteworthy that in an urban environment, the majority of priority areas are reflective (car, road, etc.).

In essence, polarimetric modality is capable of characterizing geometry by itself. This capacity is inherent to an in-depth understanding of the scene (materials, pose, light, etc.). In an unconstrained environment like urban areas, the lack of this information can cause most estimation methods to fail. Our approach proposes overcoming these constraints by using the abstraction capabilities of a deep learning network. Consequently, our method does not require a deep understanding of the scene and is therefore usable in uncontrolled urban areas.

Monodepth. There is an increased interest in the study of monocular depth estimation methods, since they offer a more flexible acquisition system as opposed to stereo systems. Most of the works are based on deep learning and optimize specific cost functions. We propose outlining some key methods that either express the basics for the domain, or similarly to our approach, impose unique constraints to the cost function refining the estimated maps.

Recent approaches alternate between training with a stereo-vision pair [23]–[25] and a monocular one [13], [26], [27]. A frequent feature of these algorithms is they are principally based on a cost function including two terms: reconstruction and smoothing. Most of the current improvements propose introducing new constraints to refine the results. [28] shows the contribution of a stereo prior cue allows an improvement in accuracy. Motion estimation has also proved its value when refining a depth map [29]. Furthermore, many other features like semantics [30], structured light pattern [31] or continuous Conditional Random Fields [32] seem to improve the estimation on an ad-hoc basis. Ultimately, the majority of the works are based on photometric features and does not transfer the achievements in non-visible image spectrum.

Our self-supervised approach introduces a possibility to append other physics-to-geometry based constraints. To be specific, we address the problem of adding polarization cues in the reconstruction cost function. Including an understanding of more complex scenes therefore reports an improving result especially on specular surfaces.

III. Method

The proposed approach is primarily based on the works of Godard et al. [13] and Berger et al. [9]. In this section, we will describe our approach constraining the loss proposed in [13] by considering polarization cues. We will subsequently discuss the methods used to train the network as well as the constraints related to the modality used.

Network and initial loss. Our approach is based on an architecture similar to the one proposed in [13] since the formulation of the loss implies geometric constraints, partly explaining the accuracy of the inference. Moreover, the key strengths of such network remain the self-supervision as well as the subdivision of the task into two parts, which makes the procedure highly adaptable. The method is self-supervised thanks to the loss which by definition does not require any ground truth. The comparison metric used in back propagation is evaluated through network output and input. The geometrically consistent loss is expressed according to a photometric reprojection error and a smoothness term, which will be discussed farther below.

Basics of polarimety. Polarimetry represents a modality sensitive to the interaction of light with the environment, thus allowing the characterization of both specular and diffuse areas of a scene.

Polarimetric information is conventionally indexed by Stokes parameters S depending on the polarizers intensities oriented at different angles $P_{\{0,45,90,135\}}$:

$$S = \begin{pmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \end{pmatrix} = \begin{pmatrix} P_0 + P_{90} \\ P_0 - P_{90} \\ P_{45} - P_{135} \\ 0 \end{pmatrix}, \tag{1}$$

with s_3 remaining null since the circular polarization is not acquired. From these Stokes parameters, one can derive

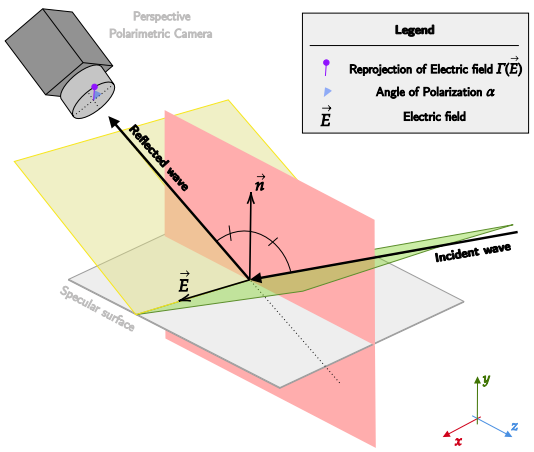


Fig. 2. Illustration of imbrication of polarimetry peculiar and perspective geometry. Visual representation of angle of polarization measurement.

three informative images characteristic of polarimetry [33], polarization angle α, degree of polarization ρ and intensity ι:

$$\begin{cases} \rho = \sqrt{(\frac{s_1}{s_0})^2 + (\frac{s_2}{s_0})^2} \\ \alpha = \frac{1}{2}\, arctan2(s_1, s_2) \\ \iota = \frac{P_0+P_{45}+P_{90}+P_{135}}{2} \end{cases} . \tag{2}$$

The intensity is then scalar, while the degree of polarization is analogous to the strength of the polarization. As shown in Figure 2, it is equally possible to geometrically represent some components of polarimetric information mixing peculiar and perspective geometry. The polarization angle α results from the angle between the reference and the direction of the electric field projection.

Prior polarimetric reconstruction error. The paper by Berger et al. [9] proposes an approach to minimize an error specific to polarimetric induced geometry. Drawing on the terms provided by Woodford et al. [18], the method consists in including a minimizable expression compelling a normal/polarization angle consistency. It is consequently shown that constraining a cost function involving a polarimetry-specific geometry is valid. Furthermore, this minimization approach is operable when optimizing a deep learning model since it depends on both the input and output of the processing pipeline and therefore could guarantees a self-supervision capability. Nevertheless, the acquisition setup as well as the problem formulation highly influence the error calculation. Indeed, [9] proposed an azimuth to acquired angle of polarization comparison. This approach is consistent under peculiar conditions implying restricted calibration of the camera or azimuth to angle of polarization specific link hypothesis. For this reason, our method proposes an alternative but similar approach allowing standard calibration and a generalized loss term releasing the constraints and allowing for easier use in real word applications.

Constraining the loss. Fundamentally, the function to minimize includes a reprojection error term and a smoothing term. Our method P2D employs the photometric error proposed in [13] since it has demonstrated its optimization and efficient convergence capabilities via deep learning. As a replacement for the edge aware first order smoothness, the second order derivative enhancement proposed by [18] is used to encourage fine transitions counterbalancing the discontinuities induced by the polarization parameters.
First, one penalizes the photometric reprojection error:

$$L_r = \min_{t'} pe(I_t, I_{t'\to t}), \tag{3}$$

with $t' \to t$ the pose transformation between two consecutive views and pe the reconstruction error:

$$pe(I_a, I_b) = \frac{\beta}{2}(1 - SSIM(I_a, I_b)) + (1-\beta)||I_a - I_b||_1. \tag{4}$$

To comply with the specifications of a minimizable function, the reprojection error comprises the weighted combination of structural dissimilarity (DSSIM) and L1 difference penalizing the deviation per pixel of the reprojection. As described in the original paper, $\beta = 0.85$ is used.

In a second step, a smoothing term is used to encourage a precise estimation of the planes while taking into account the edges:

$$L_s = |\delta_x^2 d_t^*| e^{-|\delta_x^2 I_t|} + |\delta_y^2 d_t^*| e^{-|\delta_y^2 I_t|}, \tag{5}$$

where $d_t^* = d_t/\bar{d}_t$ is the mean-normalized inverse depth enforcing the depth to be dense while reconstructing the planes [34] and the δ^2 operator is defined according to the second order prior smoothness term $\mathcal{S}(\{j,k,l\})$ [18]:

$$\mathcal{S}(\{j,k,l\}, d_t^*)_x = \delta_x^2 d_t^* = d_t^*(j) - 2 * d_t^*(k) + d_t^*(l), \tag{6}$$

with $\{j,k,l\}$ three neighboring pixels in the horizontal or vertical direction following the x-axis or y-axis orientation of the smoothing.

The weighted combination L_{diff} of these two terms then allows for a precise reconstruction of the non-reflective (diffuse) areas.

$$L_{\text{diff}} = \mu L_r + \lambda L_s, \tag{7}$$

with λ a scaling parameter set to $1e^{-3}$ and μ a binary mask defined in [13] taking occlusion and displacement of pixels along sequences into account.

Now, by drawing inspiration from and generalizing the contribution in [9], it is possible to include a third term into this loss to penalize poor reconstruction of reflective areas. By definition, polarization is defined by the orientation of the electric field. Consequently, it is possible to estimate the electric field as a function of a normal derived from a plane.

Let us consider three neighboring pixels $\{p,q,r\}$ in the layout presented in the Figure 3. This arrangement is organized

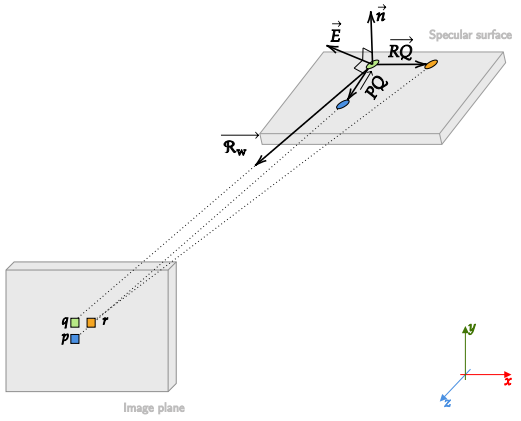


Fig. 3. Illustration of the electric field estimation method.

such that it removes fronto-parallel planes related uncertainties. Then, the projection of these three adjacent pixels into 3D results in three points of a plane, respectively $\{P, Q, R\}$. The local normal $\overrightarrow{n}$ is obtained via the cross-product of the two vectors $\overrightarrow{PQ}$ and $\overrightarrow{RQ}$ linking the points $\{P, Q, R\}$. By definition, the electric field $\overrightarrow{E(Q)}$ is perpendicular to the plane defined by the the normal and the reflected wave when considering specular surfaces. Following the definition, $\overrightarrow{E(Q)}$ at 3D point Q can be deduced from the cross product between the local normal and $\overrightarrow{\mathcal{R}_w}$ at the point Q as follows:

$$\overrightarrow{E(Q)} = \Big[\big(\Pi(p, D(p)) - \Pi(q, D(q))\big) \times \big(\Pi(r, D(r)) - \Pi(q, D(q))\big)\Big] \times \overrightarrow{\mathcal{R}_w}, \tag{8}$$

where $\Pi(x, D(x))$ is the 3D projection of pixel x relative to the disparity $D(x)$. In an optimal context, the polarization angle and the electric field maintain the same orientation and by extension the same angle relative to the reference as shown in the Figure 4. Conversely, when a depth map is incorrectly estimated, then the estimated local normal is inconsistent and consequently is the deduced polarization angle. Accordingly, we can add a term C_{pol} to the loss penalizing the deviation of the normal.

As shown in equation 9 and in Figure 4, to evaluate the deviation, it is necessary to back project the direction of the electric field onto the image plane and compare it with the angle of polarization α:

$$C_{\text{pol}}(q) = \rho(q) \left| \tan\left[\tan^{-1}\left(\Gamma(\overrightarrow{E(Q)})\right) - \alpha(q)\right]\right|, \tag{9}$$

where Γ is the back projection operator onto the image plane. Moreover, ρ allows the scaling of the loss reinforcing the necessity for correlation between α and ρ.

Since an angular differences is considered, and because this term will be combined with the reprojection term, the definition domains must be taken into account. The reprojection term

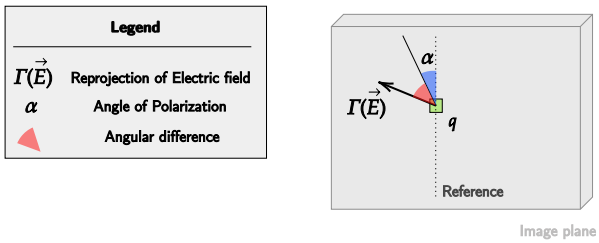


Fig. 4. Angular difference visual representation. Here, the reference of the angle of polarization is vertical.

clearly belongs to $[0, \infty[$ interval. To constrain the polarization term to the same interval, the absolute tangent is employed. As a result, the polarimetric loss term becomes:

$$L_{\text{pol}} = \frac{1}{N} \sum_{x \in \chi} C_{\text{pol}}(x), \tag{10}$$

with N the number of pixels x in the set of reference image pixels χ. Finally, the loss used to train the network is defined by:

$$\Lambda = L_{\text{diff}} + \tau\ L_{\text{pol}}, \tag{11}$$

where τ is a binary mask derived from ρ such that:

$$\tau(x) = \begin{cases} 1, & \text{if } \rho(x) \geq 0.4 \\ 0, & \text{otherwise} \end{cases}. \tag{12}$$

The polarimetric term L_{pol} is taken into account only if the degree of polarization is relevant. Since, the relevance of both ρ and α are correlated, this mask ensure for a legitimate electric field estimation. The final loss Λ is then just composed of reprojection error when the image area is unpolarized. Polarization components, when consistent, are taken into consideration and penalize the inaccurate reconstruction of specular surfaces.

Final architecture. Following [13], the network has an encoder-decoder architecture (a UNet with a ResNet 50 layout as shown in the Figure 5). It takes as input three-channel images obtained by concatenation of the intensity ι, the polarization angle α and the degree of polarization ρ. To overcome some inconsistencies related to the polarimetric modality and to consider exclusively areas with a minimum partial specularity, all values of lower than 0.4 are eliminated. This is justified by the fact that diffuse surfaces corresponding to low degree of polarization lead to a difference of $\pi/2$ between α and the electric field E. When the disparity induced by the reprojection error is calculated, despite the accuracy of this calculation, the angular error will then tend towards $\pi/2$ leading the L_{pol} function to tend towards infinity and thus causing exploding gradient problems.
Similarly, a perfect ρ is physically unobservable which justifies an upper threshold. To combine a scale-factor effect and a

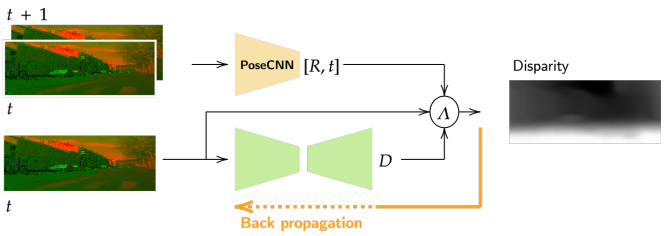


Fig. 5. Illustration of the network as well as the loss calculation strategy and its back propagation. Drawing inspiration from [13], the depth estimation network is a UNet with a ResNet50 layout.

regularization relative to physical property, the values are clipped to a maximum of 0.8.

IV. Experiments

A. Implementation details

Datasets. The training dataset was acquired during both dry and rainy weather such that the experiments would highlight the capacity of polarimetric modality in diverse conditions. All acquisitions were made with an affordable polarimetric camera, the Basler Ace aca2440-75um POL, consisting of a Sony IMX250MZR sensor delivering a resolution of 2448 x 2048 pixels. The camera was mounted on board a driving car, recording a total of approximately 7,000 images per weather condition. The final training dataset is composed of 13,400 images. As for the evaluation dataset, it is composed of a completely independent set of 25 images, acquired separately from another view under mixed meteorological conditions.

Ground truth. Ground truth generation represents a critical point when it comes to addressing urban reconstruction problems. Because of specularity, accurate depth evaluation is difficult since ground truth generation commonly rely on LiDAR sensors which are occasionally unreliable for measuring specular surfaces geometry due to reflection or transparency. Indeed, it would be a prerequisite to spray matte coating over all the specular surfaces of a complex urban scene wich is obviously unfeasible. To overcome such difficulties, the reference disparity has been pre-calculated using SGBM [35] and then refined by hand. It would have been possible to calculate the ground truth using a learning-based method. It should be considered the approach presented here is to improve deep learning methods since they typically fail on specular surfaces. Moreover, it is notable the vast majority of networks are trained on the same database which consists of images in favorable weather conditions. For these reasons, the choice of a refined SGBM eliminates learning biases while providing ground truth taking into account specular surfaces. This approach is unconventional but permits to conceive a global idea of the reliability of the results while allowing the computation of metrics. In addition, the disparity remains a relative value and therefore the impact of manual refinement is minor.

Network training. The network was trained on a machine consisting of a Nvidia Titan Xp (12GB memory) GPU, 128GB of RAM and two CPUs accumulating a total of 24 physical cores. We use the following parameters for all the networks: a batch size of 12, a learning rate of 1e-3 and a maximum of 30 epochs. For a fast training, the images were downsampled without any interpolation method to maintain the physical properties. Following this routine, training with polarimetric images takes approximately 17 hours compared to 12 hours when training with intensity images only. The forward pass inference time is around 0.45 second per image in pure CPU processing.

Evaluation. We compared the results of our method P2D with the competitive state-of-the-art method described in [13]. Our P2D receives as input the polarization parameters by concatenation of the three channels $\{\iota, \alpha, \rho\}$. For the method in [13], we evaluate two versions. One version, G_{RGB}, using only intensity images and trained with the weights provided by the authors without fine-tuning. And, another version, G_{I}, trained in an end-to-end manner so that the network parameters are adapted to the intensity images at hand.

Metrics. The calculated metrics shown in the table I represents popular assessments within the reconstruction community that have been proposed by Eigen et al. [36]. They provide an unbiased and comprehensive measure of results. In particular, the δ values are calculated on the prediction/ground truth ratio and highlight an intrinsic precision of the reconstruction.

At last, the sky reconstruction accuracy R_s is calculated as follows:

$$R_s = 1 - \left(\frac{\hat{y}_s}{y_s}\right), \tag{13}$$

where y_s is the sum of the binary masked pixels considered as sky in the ground truth and $\hat{y}_s$ the corresponding area in the prediction. This calculation is performed on the disparity, and one order of magnitude error deviation is considered acceptable. It focuses on the ability of the network to accurately estimate the sky and not propagate an erroneous evaluation in such areas. It is noteworthy this kind of precision is usually neglected since the reconstruction precision of these areas is removed from the frequent metrics. Ordinarily, sky zones are filtered out of the metrics beforehand. In this contribution, these areas are also neglected while calculating Eigen et al. [36] metrics.

B. Results and discussion

Table I and Figure 6 allow for a quantitative and qualitative evaluation of the results. Analyzing the images in Figure 6, we can observe various responses of the networks. First, using the method in [13] with raw images (G_{RGB}), the results seem satisfactory at first glance. However, some characteristics of

TABLE I

QUANTITATIVE COMPARATIVE RESULTS. FOR EACH NETWORK SEVERAL METRICS ARE COMPUTED NEGLECTING THE SKY AREAS. IN ADDITION, WE PROPOSE THREE DIFFERENT EVALUATIONS: ON THE *Raw* IMAGES AT THE OUTPUT OF THE NETWORK, ON THE *Cropped* IMAGES TO ELIMINATE INCONSISTENCIES IN THE POLARIMETRIC NETWORK, AND ON THE *Specular* AREAS ONLY. G_{RGB} CORRESPONDS TO THE NETWORK PRESENTED IN [13] WITHOUT FINE-TUNING, AND G_I CORRESPONDS TO THE SAME NETWORK WITH FINE-TUNING. P2D CORRESPONDS TO OUR METHOD.

Type	**Network**	Abs_Rel	Sq_Rel	RMSE	RMSE_log	$\delta > 1.25$	$\delta > 1.25^2$	$\delta > 1.25^3$
Raw	G_{RGB}	0.471	10.809	25.161	0.680	0.485	0.707	0.804
	G_I	0.482	9.144	22.332	0.617	0.431	0.695	0.838
	P2D	**0.322**	**4.504**	**20.651**	**0.484**	**0.537**	**0.801**	**0.896**
Cropped	G_{RGB}	0.533	14.050	29.312	0.780	0.449	0.658	0.771
	G_I	0.415	11.247	25.899	0.678	0.467	0.729	0.850
	P2D	**0.245**	**5.650**	**24.009**	**0.531**	**0.604**	**0.825**	**0.910**
Specular	G_{RGB}	0.341	8.249	7.236	0.306	0.666	0.808	0.896
	G_I	0.208	2.248	5.491	0.233	0.639	0.877	0.952
	P2D	**0.147**	**1.583**	**4.898**	**0.166**	**0.796**	**0.921**	**0.973**

the images are altered. For example, specular areas, car windshields or bus stops are incorrectly detected. To be specific, car windshields are over-segmented into several parts rather than being detected as unique planar surface. In addition, the distance to reflective road lines is often under-estimated and farthest objects are ignored. However, as the weights of the network are not fine-tuned, its features representations have been learned exclusively from textures characteristics which limit the performance of the method in specular or reflective areas. Nevertheless, the reconstruction is close enough to the ground truth which also shows the robustness of this approach and reinforces the initial idea of using it as a baseline method.

When the network is trained end-to-end with polarimetric intensity images, the global impact of polarization is reduced but brings many significant constraints. Despite an accurate estimation of some areas like planar surfaces on cars and long distance objects (see row G_I in Figure 6), others areas are subject to some aberrations mainly on the reflective lines and polarized contours. This behaviour produces a direct impact on the network estimations. Hence, the addition of polarization-specific terms is necessary to improve the estimation of polarized areas.

When employing all the polarimetric information (ι, ρ and α) in our P2D network, we can observe a accurater estimation of specular areas as well as sufficient reconstruction of diffuse areas.
We can however see that the results at limited distances, as shown in the P2D row of Figure 6, are occasionally incorrect. This is due to the fact polarimetric information varies according to the light and its reflection angle. Therefore, the position of the camera is primarily responsible for these erroneous estimates. Indeed, as explained earlier in the Datasets section, the images of the evaluation subsets were acquired with different camera poses. Consequently, the information from the images differs from the training set case leading the network to fail in estimating depth values at close distances. To have an estimation in favorable conditions, the choice of cropping the lower quarter of the images for a second evaluation is proposed. Note, this lower part corresponds to closer distances. Both the estimates and the ground truth depth maps are cropped. These results are shown in the $P2D_C$ line of Figure 6 as well as the *Cropped* part of the Table I. We can observe better performances especially when comparing errors with the δ metric. The most considerable improvements are obtained when looking at both $\delta > 1.25$ and $\delta > 1.25^2$ showing a respective 16% and 17% improvement compared to the G_{RGB} network.

Additionally, we perform an evaluation considering only the specular areas of the scenes. This is achieved using a rule-based naive system filtering polarization degrees higher than 0.4, hence keeping areas which are highly specular. The results shown in the last part of Table I, exhibit improved performances for all the evaluated networks since the assessed pixel space is reduced. However, the largest improvement is obtained with our P2D method. Specifically, our method achieves 92% for $\delta > 1.25^2$ ratio, compared to 80% obtained by the state-of-the-art method. Consequently, we can see the polarimetric modality is beneficial for the reconstruction of urban scenes with many specular surfaces. Ultimately, to highlight the depth map reliability, sky reconstruction accuracy (eq. 13) has been computed in Table II.

TABLE II

QUANTITATIVE COMPARISON OF SKY RECONSTRUCTION ACCURACY.

Network	G_{RGB}	G_I	P2D
R_s	0.055	0.388	**0.532**

This specific metric has been computed since many evaluation metrics neglect such aspect which, however, could be informative, especially if one uses such an algorithm for navigation. This ratio reveals P2D's ability to reconstruct slightly more than half of the sky correctly. It permits to demonstrate polarimetry to be favorable also for such estimation.

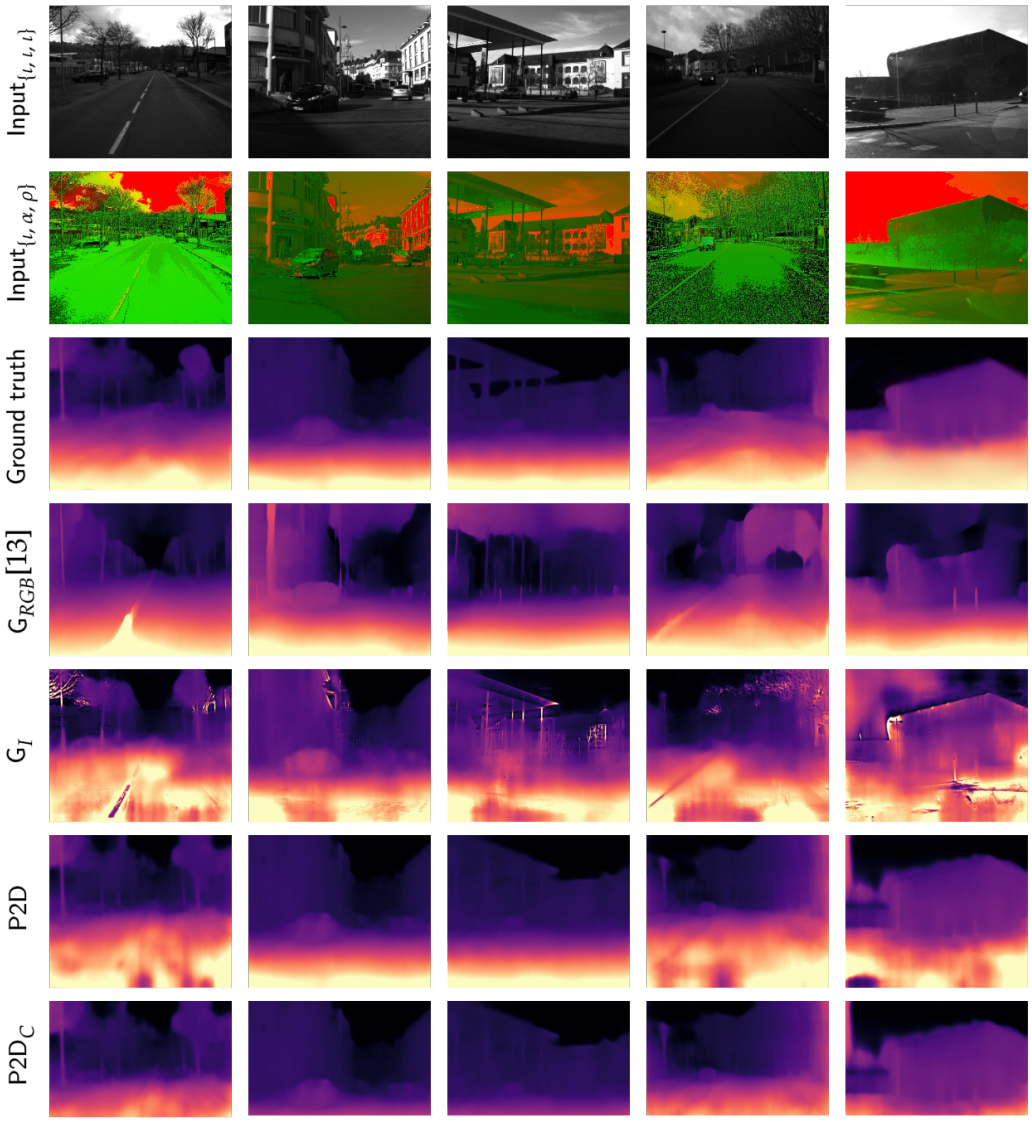


Fig. 6. Illustration of results on five independent road scenes in mixed weather conditions. From top to bottom, the inputs to the two networks (scalar or polarimetric), the ground truth depth map, then the results of the three different networks: G_{RGB} corresponds to the network presented in [13] without fine-tuning, and G_I corresponds to the same network with fine-tuning. P2D corresponds to our method. The last row shows the crop version of the results from P2D to eliminate inconsistencies due to the modality and aberrations because of the camera position. Columns one and four correspond to acquisitions in light rainy weather, hence, the different road behaviour in polarimetric space. The other images are acquired under normal conditions.

V. Conclusion

In recent years, the cost of polarimetric cameras has been on a decreasing trend. This allows the use of a more affordable alternative type of information and thus an opening to data directly related to the physics of the scenes. We have proposed a method to include polarization cues in a cost function applied to monocular depth estimation. With this polarimetry adapted loss, we can directly consider geometric constraints specific to the scene.

We next showed the proposed approach was remarkably effective for depth estimation in urban environment. Moreover, unlike previous methods, polarization allows being insensitive to specular surfaces. This is a significant advantage since traditional algorithms have difficulties to cope with such areas.

It is explicit that a fusion method using an RGB and polarimetric pair would allow an accurate and stable reconstruction in any condition. Despite this observation, one of the considerable challenges of this approach remains the multimodal/multifocal alignment. Recent sensors, like the Sony Pregius IMX250MYR-C, would potentially allow combining the strengths of each modality by merging the diverse types of information. It would be beneficial to take advantage of this modern technology to propose an accurate reconstruction of both specular and diffuse areas. Ultimately, it is noticeable

that RGB aligned images would allow a completely fair and unbiased comparison for the methods.

ACKNOWLEDGMENTS

This work was supported by the French National Research Agency through ANR ICUB (ANR-17-CE22-0011). We gratefully acknowledge the support of NVIDIA Corporation with the donation of GPUs used for this research.

REFERENCES

[1] H. Jiang, G. Larsson, M. Maire Greg Shakhnarovich, and E. Learned-Miller, "Self-supervised relative depth learning for urban scene understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 19–35.

[2] H. Zhang, C. Shen, Y. Li, Y. Cao, Y. Liu, and Y. Yan, "Exploiting temporal consistency for real-time video depth estimation," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct. 2019.

[3] J. Jiao, Y. Cao, Y. Song, and R. Lau, "Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 53–69.

[4] T. v. Dijk and G. d. Croon, "How do neural networks see depth in single images?" in *The IEEE International Conference on Computer Vision (ICCV)*, Oct. 2019.

[5] J. Hu, Y. Zhang, and T. Okatani, "Visualization of convolutional neural networks for monocular depth estimation," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct. 2019.

[6] P.-Y. Chen, A. H. Liu, Y.-C. Liu, and Y.-C. F. Wang, "Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019.

[7] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[8] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[9] K. Berger, R. Voorhies, and L. H. Matthies, "Depth from stereo polarization in specular scenes for urban robotics," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1966–1973.

[10] Z. Cui, J. Gu, B. Shi, P. Tan, and J. Kautz, "Polarimetric multi-view stereo," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] M. Rastgoo, C. Demonceaux, R. Seulin, and O. Morel, "Attitude estimation from polarimetric cameras," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems-IROS*, 2018.

[12] C. V. Nguyen, M. Milford, and R. Mahony, "3d tracking of water hazards with polarized stereo cameras," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5251–5257.

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

[14] S. Rahmann and N. Canterakis, "Reconstruction of specular surfaces using polarization imaging," vol. 1, 02 2001, pp. I–149.

[15] O. Morel, C. Stolz, F. Meriaudeau, and P. Gorria, "Active lighting applied to three-dimensional reconstruction of specular metallic surfaces by polarization imaging," *Applied optics*, vol. 45, no. 17, pp. 4062–4068, 2006.

[16] R. Blin, S. Ainouz, S. Canu, and F. Meriaudeau, "A new multimodal rgb and polarimetric image dataset for road scenes analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 216–217.

[17] A. E. R. Shabayek, C. Demonceaux, O. Morel, and D. Fofi, "Vision based uav attitude estimation: Progress and insights," *Journal of Intelligent & Robotic Systems*, vol. 65, no. 1-4, pp. 295–308, 2012.

[18] O. J. Woodford, "Global Stereo Reconstruction under Second Order Smoothness Priors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 12, pp. 2115–2128, 2008.

[19] L. Yang, F. Tan, A. Li, Z. Cui, Y. Furukawa, and P. Tan, "Polarimetric dense monocular slam," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3857–3866.

[20] A. Kadambi, V. Taamazyan, B. Shi, and R. Raskar, "Polarized 3d: High-quality depth sensing with polarization cues," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3370–3378.

[21] D. Zhu and W. A. P. Smith, "Depth from a polarisation + RGB stereo pair," *CoRR*, vol. abs/1903.12061, 2019. [Online]. Available: http://arxiv.org/abs/1903.12061

[22] J. Keller and S. Scherer, "A stereo algorithm for thin obstacles and reflective objects," *arXiv preprint arXiv:1910.04874*, 2019.

[23] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 340–349.

[24] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia, "Every pixel counts: Unsupervised geometry learning with holistic 3d motion understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[25] 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.

[26] 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*, 2017, pp. 270–279.

[27] 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 Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 240–12 249.

[28] J. Watson, M. Firman, G. J. Brostow, and D. Turmukhambetov, "Self-supervised monocular depth hints," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2162–2171.

[29] F. Brickwedde, S. Abraham, and R. Mester, "Mono-sf: Multi-view geometry meets single-view depth for monocular scene flow estimation of dynamic traffic scenes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2780–2790.

[30] P.-Y. Chen, A. H. Liu, Y.-C. Liu, and Y.-C. F. Wang, "Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2624–2632.

[31] G. Riegler, Y. Liao, S. Donne, V. Koltun, and A. Geiger, "Connecting the dots: Learning representations for active monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7624–7633.

[32] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5354–5362.

[33] E. Collett, "Field guide to polarization." Spie Bellingham, WA, 2005.

[34] C. Wang, J. Miguel 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*, 2018, pp. 2022–2030.

[35] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–341, 2007.

[36] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, 2014, pp. 2366–2374.