ON MONOCULAR DEPTH ESTIMATION AND UNCERTAINTY QUANTIFICATION USING CLASSIFICATION APPROACHES FOR REGRESSION

Xuanlong Yu¹,²      Gianni Franchi²      Emanuel Aldea¹

¹SATIE, Paris-Saclay University      ²U2IS, ENSTA Paris, Institut Polytechnique de Paris

ABSTRACT
Monocular depth is important in many tasks, such as 3D reconstruction and autonomous driving. Deep learning based models achieve state-of-the-art performance in this field. A set of novel approaches for estimating monocular depth consists of transforming the regression task into a classification one. However, there is a lack of detailed descriptions and comparisons for Classification Approaches for Regression (CAR) in the community and no in-depth exploration of their potential for uncertainty estimation. To this end, this paper will introduce a taxonomy and summary of CAR approaches, a new uncertainty estimation solution for CAR, and a set of experiments on depth accuracy and uncertainty quantification for CAR-based models on KITTI dataset. The experiments reflect the differences in the portability of various CAR methods on two backbones. Meanwhile, the newly proposed method for uncertainty estimation can outperform the ensemble method with only one forward propagation.

Index Terms— Depth estimation, Uncertainty Estimation

1. INTRODUCTION
In machine learning, regression tasks predict a continuous output based on a given input. Yet, if the ground truth (prediction target) is within a specific range, e.g., in the case of age estimation [1], one can quantize the ground truth and cast regression into a classification problem. We refer to these techniques as Classification Approaches for Regression (CAR), and in this paper we explore CAR techniques applied on monocular depth estimation.

Monocular depth estimation (MDE), which is an ill-posed problem [2], consists in predicting the scene depth given only an RGB image as the input. Deep Neural Networks (DNNs) learn the mapping between the single RGB images and their corresponding depth maps to solve MDE, and show good performance on indoor and outdoor benchmarks [3, 4].

Classification Approaches for Regression (CAR) [5, 6, 7, 8, 9, 10] have emerged recently in the spotlight among MDE algorithms. The core idea is to transfer regression to a classification problem using quantization (or discretization) strategies. The classification models can natively provide the confidence for prediction results, which also has the potential to improve the prediction accuracy [8]. DNNs are prone to two kinds of uncertainty: aleatoric uncertainty and epistemic uncertainty [11]. It is crucial to study the uncertainty of DNNs if we want to rely on their predictions. Some works proposed to estimate the uncertainty of MDE DNNs by using an auxiliary network [12, 13], or ensembling [14]. Here we gain access to the uncertainty directly using the CAR DNN.

This work will investigate and show the complete picture of the CAR MDE methods. The contributions are as follows:
1. We systematically summarize and formalize the all major CAR MDE mechanisms to the best of our knowledge;
2. We implement these mechanisms on top of two different backbones, comparing depth prediction and uncertainty quality on various evaluation metrics;
3. We propose a new, effective uncertainty estimation method named Expectation of Distance for CAR MDE models.

2. OVERVIEW OF CLASSIFICATION APPROACHES FOR REGRESSION (CAR)
2.1. Taxonomy of CAR MDE
To better unify the terms and make it easier to grasp the differences in contributions of CAR strategies, we propose to decompose the CAR problems into three key components: discretization, loss function and post-processing. Table 1 offers an overview of the specific strategies used in the previous works. The details are provided in the following sections.

The contributions of previous CAR based MDE solutions fall in two main groups: 1. novel strategies in the three components mentioned above [5, 6, 7, 8, 9, 10]; 2. architecture and/or loss modifications based on the previous strategies [16, 19, 17, 20, 15, 18]. In most papers, CAR can improve model accuracy, making it outperform its regression version [5, 6, 7, 8, 9, 10, 16], or may improve model per-

We acknowledge the support of the Saclay-IA computing platform.
formance as part of multi-task learning [18, 17].

2.1.1. General notations

Let us first consider a monocular depth dataset \( D = \{ (x_i, d_i) \} \), where \( x_i \in \mathbb{R}^{3 \times H \times W} \) and \( d_i \in (\mathbb{R}^+)^{H \times W} \) represents the ground truth depth \( d_i \) for the image \( x_i \). We denote \( \{ d_{i,j} \}_{j} \) all the pixel values in \( d_i \), where \( N \) is the number of pixels with valid ground truth. \( a, b \) are two real values representing the minimum and maximum depth value for the dataset.

For CAR strategies, we denote \( K \) the number of classes, which represents the level of discretization. Additionally, to simplify the notations, we use \( \log \) as the logarithm with base \( e \) for all papers except for [16, 18], where the \( \log \) refers to the logarithm with base 10. We denote \( f_{\theta_1} \) the DNN with parameters \( \theta_1 \). Given \( x_i \), the prediction of \( f_{\theta_1} \):

\[
y_i = \frac{e^{f_{\theta_1}(x_i)}}{\sum_{p=1}^{K} e^{f_{\theta_1}(x_i)_p}} \tag{1}
\]

where \( f_{\theta_1}(x_i) \in \mathbb{R}^{c \times H \times W} \) is the logit map and \( [f_{\theta_1}(x_i)_p] \) its \( p \)-th element, and \( \gamma_i \) is the Softmax output. The number of its channels \( c = K \) by default, otherwise equals to the specific settings as in DORN [7] (2K) and Adabins [10] (128).

2.1.2. Discretization: Fully Handcrafted

The discretization function will output two components given \( d \): a depth table \( \hat{d} = \{ d_{p} \}_{p=1}^{K} \in \mathbb{R}^{K} \) where the possible discrete depth values are set ordinally, and an indicator map equivalent to a classification map \( \gamma_i = \{ [y_{i,j,p}]_{p=1}^{K} \}_{j} \in (\mathbb{R}^+)^{K \times H \times W} \) which points for each pixel the closest discrete depth value. This closest depth value can be considered as a class, leading to a classification task. Both \( \hat{d} \) and \( \gamma_i \) are handcrafted, and the goal becomes the learning of \( \gamma_i \).

Handcrafted \( \hat{d} \): \( \hat{d} \) contains \( K \) values representing the centers of intervals \([\hat{d}_{0}, \hat{d}_{1}, ..., \hat{d}_{K-1}, \hat{d}_{K}]\) with an interval width \( q \):

\[
\hat{d} = \{ d_{p} \}_{p=1}^{K} = \{(\hat{d}_{0} + \hat{d}_{1}/2, ... , (\hat{d}_{K-1} + \hat{d}_{K})/2) \} \tag{2}
\]

Handcrafted \( \hat{d} \) + One-hot \( \gamma_i \): Given \( \hat{d} \), constructing \( \gamma_i \) is done using one-hot encoding, as applied in [6, 16, 17, 18]:

\[
y_{i,j,0}^{[\text{onehot}]} = \{ y_{i,j,0}, \ldots, y_{i,j,k}, \ldots, y_{i,j,K-1} \} \in \mathbb{R}^{K} \tag{3}
\]

with \( y_{i,j,0} = 1 \), if \( k = \lfloor \log(d_{i,j})/q \rfloor \) and 0 otherwise.

in which, \( \lfloor \cdot \rfloor \) is a rounding operator, \( q \) is defined in Eq. 2.

Handcrafted \( \hat{d} \) + Ordinal \( \gamma_i \): Furthermore, there are several variants of Eq 3. Ordinal properties can be applied as it is presented in [7] and the followed works [19, 20]:

\[
y_{i,j}^{[\text{ord}]} = \{ y_{i,j,0}, \ldots, y_{i,j,k}, \ldots, y_{i,j,K-1} \} \in \mathbb{R}^{K} \tag{4}
\]

with \( y_{i,j,0} = 1 \), if \( k = \lfloor \log(d_{i,j})/a \rfloor /q \), and 0 otherwise.

Handcrafted \( \hat{d} \) + Smooth \( \gamma_i \): It is also possible to predict a smooth discrete map from the initial discrete map \( \hat{y}_{i,j}^{[\text{onehot}]} \). The indicator in the classification map is softened by applying on \( \hat{d} \) a Gaussian kernel, as to predict distance within a coarser range. The smooth \( \gamma_i \) are defined by:

\[
y_{i,j}^{[\text{smo1}]} = e^{-\gamma\text{||} \log(d_{i,j}) - \hat{d} \text{||}^2} \tag{5}
\]

\[
y_{i,j}^{[\text{smo2}]} = e^{-\gamma\text{||} \log(d_{i,j}) - \hat{d} \text{||}^2} \sum_{p=0}^{K-1} e^{-\gamma\text{||} \log(d_{i,j}) - d_{p} \text{||}^2} \tag{6}
\]

where \( \gamma \) is a hyperparameter which can be regarded as the scale of the discrete distribution (the smaller \( \gamma \), the flatter the label distribution in \( y_{i,j} \)). Specifically, Yang et al. [9] use Eq 5 as the unnormalized soft target labels, while SORN [8] applies the normalized version in Eq 6. Moreover, Cao et al. [5] introduce a \( K \times K \) symmetric “information gain” matrix \( H \) in their loss function with elements \( H(k,p) = e^{-\gamma\text{||} k-p \text{||}^2} \), where \( p \in \{0, ..., K-1\} \) and \( k \) is the discrete ground truth index as defined in Eq 3. In this case:

\[
y_{i,j}^{[\text{smo3}]} = e^{-\gamma\text{||}k-p\text{||}^2} = e^{-\gamma q^{-2}\text{||} \log(d_{i,j}) - (\hat{d} - 0.5q) \text{||}^2} \tag{7}
\]

Since \( q \) is a constant, this strategy can be regarded as being equivalent to the \( \hat{y}_{i,j}^{[\text{smo1}]} \) in Eq 5.

2.1.3. Discretization: Adaptive

In the absence of the handcrafted depth table or the classification map, one may also implicitly train both of them using a regression loss as in Adabins [10]. Thus the goal of the DNN is changed from fitting the handcrafted classification maps to fitting the continuous ground truth depth, while still following the principle of building depth tables and classification maps.

In Adabins [10], the depth table is implicitly trained along with the classification map using a non-linear block \( g_{\theta_2} \) with parameters \( \theta_2 \) of \( g \), which is a mini ViT [10]. In this case, \( g_{\theta_2} \) is set on top of the backbone \( f_{\theta_1} \), and it will output \( \hat{d}_{i,p}^{[\text{ada}]} \) and \( \hat{y}_{i,j}^{[\text{ada}]} \) given \( f_{\theta_1}(x_i) \):

\[
\hat{d}_{i,p}^{[\text{ada}]} = \{ a - \alpha/(\sum_{s=0}^{p} \hat{d}_{i,s}^{[\text{ada}]})) \}_{p=0}^{K-1} ; \hat{y}_{i,j}^{[\text{ada}]} = \frac{e^{\hat{y}_{i,j}}}{\sum_{s=0}^{K-1} e^{\hat{y}_{i,j,s}}} \tag{8}
\]

where \( \gamma_{i,j}^{[\text{ada}]} \) is the product of a Softmax function, and \( \hat{d}_{i,p}^{[\text{ada}]} \) is a cumulative summation output followed by a normalization operation which is included in \( g_{\theta_2} \). Since \( \hat{d}_{i,p}^{[\text{ada}]} \) is a product of \( g_{\theta_2} \) taking \( f_{\theta_1}(x_i) \), for each \( x_i \), not only a unique classification map but also a unique depth table will be provided.

2.1.4. Loss function

Based on the previous discretization strategies, we introduce here the loss function design. Models should fit their output to the designed \( y \) or \( \hat{d} \). For brevity, we define first the total loss \( L_{\text{total}} = \sum_{i}^{N} \sum_{j=0}^{N-1} L_{i,j} \) where \( L_{i,j} \) is the loss for pixel \( j \), on the \( i \)-th data. We just define \( L_{i,j} \) in the following sections for simplicity.

Cross entropy (CE) loss: is a straightforward solution given an one-hot classification map: \( L_{i,j}^{\text{CE}}(\theta_1) = -(y_{i,j}^{[\text{onehot}]} \text{log} \gamma_{i,j}) \).
**Ordinal regression loss**: is essentially an implicit ordinal selection plus a multiple binary cross entropy (BCE) loss. Instead of directly using $y_{i,j}$, it requires to do an ordinal selection on the logit map $f_{i,j}(x_i)$ with $c = 2K$ to $c = K$ as the predicted classification map, then to apply a Multiple-BCE loss on it:

$$L_{i,j}^{[ord]}(\theta_1) = -[y_{i,j}^{[ord]} \log y_{i,j}^{[ord]} + (1 - y_{i,j}^{[ord]}) \log(1 - y_{i,j}^{[ord]})]$$

(9)

with $y_{i,j}^{[ord]} = \frac{e^{f_{i,j}(x_i)_2p+1}}{e^{f_{i,j}(x_i)_2p+1} + \sum_{p=0}^{K-1} e^{f_{i,j}(x_i)_2p}}$

where $2p + 1$ and $2p$ represent the indices of the coefficient.

**Weighted CE loss**: is applied when the target vector is a soft discrete distribution. The CE loss turns to be equal to: $L_{i,j}^{[WCE]}(\theta_1) = -[y_{i,j}^{[smo]} \log y_{i,j}^{[smo]}]$, and it has the same form as the Kullback-Leibler divergence loss.

**Multiple BCE loss**: is another solution when the target is a soft discrete distribution. Yang et al. [9] apply BCE loss on every class value in $y_{i,j}^{[smo]}$ defined in Eq. 5.

**Regression losses**: DS-SIDENet [18] applies CAR with a smooth L1 loss [21] to fit the one-hot classification map target. Conversely, Adabins [10] combines the per-image adaptive depth table $d_{i,j}^{[ada]}$ and $y_{i,j}^{[ada]}$ defined in Eq. 8 as the predicted depth, then applies a Scale-Invariant loss [22].

### 2.1.5 Post-processing

Post-processing aims to restore the discrete predicted labels to continuous depth values. In the following sections, we use the power function in base $e$, see Sec. 2.1.1.

**Ordinal sum**: For DORN [7], the continuous depth is restored from the sum of the output Sigmoid labels which are higher than or equal to 0.5:

$$d_{i,j} = \exp[\log(a) + q \cdot \sum_{p=0}^{K-1} 1\{y_{i,j,p}^{[ord]} \geq 0.5\} + 0.5]$$

(10)

where $q$ and $y_{i,j,p}^{[ord]}$ are defined in Eq. 2 and Eq. 9 respectively.

**Soft weighted sum**: is a solution that may applied on both handcrafted or learned depth tables. It sums the Hadamard product between the depth table and the classification table:

$$d_{i,j} = \exp\{\sum_{p=0}^{K-1} d_p \cdot y_{i,j,p}\}$$

(11)

Note that essentially AdaBins [10] also follows this pattern.

**Argmax**: The authors of SORN [8] claim that Argmax outperforms Soft weighted sum in their case:

$$d_{i,j} = \exp[\log(a) + q \cdot \max_p\{y_{i,j,p}\}^{K}] + 0.5$$

(12)

where $q$ is defined in Eq. 2.

### 2.2 Uncertainty estimation of CAR MDE

In this section, we will discuss the previous works on uncertainty estimation for CAR MDE, the difficulty of this problem and our proposed approaches on estimating CAR uncertainty. The ground truth uncertainty or the Oracle should be the model’s prediction error. The previous works on MDE uncertainty estimation [13, 12] mainly use the principle of learning the prediction error [11]. Meanwhile, the Variance among the point estimations given by MC-Dropout [23] and Deep Ensembles [14] can also be applied for this task. Unlike the previous works, the likelihoods of the predicted class (the quantified depth value) given the input data provided by CARs can offer another possibility to estimate the uncertainty mentioned in the previous works but rarely discussed. Yang et al. [9] suggest to use Shannon Entropy (S-Entr) [24] among the output Softmax classification map: $-\sum_{p=0}^{K-1} y_{i,j,p} \log y_{i,j,p}$. Moreover, they showed cases where the depth is well predicted, yet the entropy is high, leading to an under-confident uncertainty score. Other strategies such as 1-Maximum Class Probability (1-MCP) can also be regarded as the uncertainty: $1 - \max_p\{y_{i,j,p}\}$. These are typical solutions used in classification tasks, and we argue that they will be suitable in case of using Argmax in post-processing for CAR problems. Widely used soft weighted sum (see Table 1) makes the property of CAR special: not only the classification map but also the depth table should be taken into account in the final result as shown in Eq. 11.

Following these remarks, we propose a new solution for CAR uncertainty. We first note that the previously mentioned methods lack consideration of the depth table, and further its relationship to the classification map. Hence, we define as CAR uncertainty metric: the Expectation of Distance (E-Dist) between the quantified depth values (either handcrafted logarithm depth table $\hat{d}$ (Eq. 2) or the learned one $d_{i,j}^{[ada]}$ (Eq. 8)) and the final predicted depth $\hat{d}_{i,j}$ (Eq. 11, 12):

$$\text{E-Dist} = \sum_{p=0}^{K-1} y_{i,j,p} \cdot (\hat{d}_{i,j} - d_{i,j})^2$$

or

$$\text{E-Dist} = \sum_{p=0}^{K-1} y_{i,j,p} \cdot (\hat{d}_{i,j}^{[ada]} - d_{i,j})^2$$

(13)

Additionally, to our knowledge, no previous works discuss the uncertainty of ordinal regression model [7]. According to its CAR strategy, only values greater than or equal to 0.5 in its classification map will be considered in the final depth calculation, thus we argue that the uncertainty comes from this part. The modified $\text{E-Dist}$ for ordinal regression is as below:

We propose to discretize the depth prediction $\hat{d}_{i,j}$ (Eq. 10) using Eq. 4, that we denote as $y_{i,j}^{[ord]}$. Then we calculate the distance between $y_{i,j}^{[ord]}$ and $y_{i,j}^{[ord]}$ (defined in Eq.9) weighted by the depth table and only consider the part with $y_{i,j}^{[ord]} \geq 0.5$:

$$\sum_{p=0}^{K-1} e^{d_p} \cdot (y_{i,j,p}^{[ord]} - y_{i,j,p}^{[ord]})^2 \cdot 1\{y_{i,j,p}^{[ord]} \geq 0.5\}$$

(14)

### 3. EXPERIMENTS

In this section, we fill in the missing comparisons of the previous works. Meanwhile, our experiments provide an extensive analysis of CAR MDE uncertainty estimation. While it is not trivial to propose a model-agnostic approach, the ensuing discussion establishes some important guidelines about performing this task on CAR models.
Table 2: The best/second-best values are highlighted in dark/light blue. The results from regression-based models are provided as reference (lower rows), and we only highlight the CAR-based results. (a) Depth uncertainty evaluations. MC-Dropout and Deep Ensembles will only provide the uncertainty with one forward pass) and Deep Ensembles [14] (with 3 models).

Table 3: Time consumption on Forward+Backward passes for one image.

Fig. 1: Experiment pipeline. Three heads will be applied on the backbone: I. original version; II. MDE with handcrafted discretization; III. MDE with adaptive discretization through a mini ViT module [10].
5. REFERENCES

[1] Gil Levi and Tal Hassner, “Age and gender classification using convolutional neural networks,” in CVPR workshops, 2015.

[2] Pawan Sinha and Edward Adelson, “Recovering reflectance and illumination in a world of painted polyhedra,” in ICCV, 1993.

[3] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger, “Sparsity invariant cnns,” in 3DV, 2017.

[4] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus, “Indoor segmentation and support inference from rgbd images,” in ECCV, 2012.

[5] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen, “Estimating depth from monocular images as classification using deep fully convolutional residual networks,” TCSVT, 2017.

[6] Bo Li, Yuchao Dai, and Mingyi He, “Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference,” Pattern Recognition, 2018.

[7] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao, “Deep ordinal regression network for monocular depth estimation,” in CVPR, 2018.

[8] Raul Diaz and Amit Marathe, “Soft labels for ordinal regression,” in CVPR, 2019.

[9] Gengshan Yang, Peiyun Hu, and Deva Ramanan, “Inferring distributions over depth from a single image,” in IROS, 2019.

[10] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka, “Adabins: Depth estimation using adaptive bins,” in CVPR, 2021.

[11] Alex Kendall and Yarin Gal, “What uncertainties do we need in bayesian deep learning for computer vision?,” NIPS, 2017.

[12] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia, “On the uncertainty of self-supervised monocular depth estimation,” in CVPR, 2020.

[13] Xuanlong Yu, Gianni Franchi, and Emanuel Aldea, “Slurp: Side learning uncertainty for regression problems,” in BMVC, 2021.

[14] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” in NIPS, 2017.

[15] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan, “Enforcing geometric constraints of virtual normal for depth prediction,” in ICCV, 2019.

[16] Ruibo Li, Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, and Lingxiao Hang, “Deep attention-based classification network for robust depth prediction,” in ACCV, 2018.

[17] Lukas Liebel and Marco Körner, “Multidepth: Single-image depth estimation via multi-task regression and classification,” in ITSC, 2019.

[18] Haoyu Ren, Mostafa El-Khamy, and Jungwon Lee, “Deep robust single image depth estimation neural network using scene understanding.,” in CVPR Workshops, 2019.

[19] Chen-Chou Lo and Patrick Vandewalle, “Depth estimation from monocular images and sparse radar using deep ordinal regression network,” in ICIP, 2021.

[20] Minh Hieu Phan, Son Lam Phung, and Abdelkamr Bouzerdoum, “Ordinal depth classification using region-based self-attention,” in ICPR, 2021.

[21] Jia-Ren Chang and Yong-Sheng Chen, “Pyramid stereo matching network,” in CVPR, 2018.

[22] David Eigen, Christian Puhrsch, and Rob Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in NIPS, 2014.

[23] Yarin Gal and Zoubin Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in ICML, 2016.

[24] Claude Elwood Shannon, “A mathematical theory of communication,” ACM SIGMOBILE MC2R, 2001.

[25] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh, “From big to small: Multi-scale local planar guidance for monocular depth estimation,” arXiv preprint arXiv:1907.10326, 2019.

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell, “Fully convolutional networks for semantic segmentation,” in CVPR, 2015.

[27] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, “Densely connected convolutional networks,” in CVPR, 2017.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in CVPR, 2016.

[29] Gianni Franchi, Andrei Bursuc, Emanuel Aldea, Severine Dubuisson, and Isabelle Bloch, “One versus all for deep neural network incertitude (ovnni) quantification,” IEEE Access, 2021.