2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

Bidirectional Attention Network for Monocular Depth Estimation

Shubhra Aich*, Jean Marie Uwabeza Vianney*, Md Amirul Islam, Mannat Kaur, and Bingbing Liu†

Abstract— **In this paper, we propose a Bidirectional Attention Network (BANet), an end-to-end framework for monocular depth estimation (MDE) that addresses the limitation of effectively integrating local and global information in convolutional neural networks. The structure of this mechanism derives from a strong conceptual foundation of neural machine translation, and presents a light-weight mechanism for adaptive control of computation similar to the dynamic nature of recurrent neural networks. We introduce bidirectional attention modules that utilize the feed-forward feature maps and incorporate the global context to filter out ambiguity. Extensive experiments reveal the high degree of capability of this bidirectional attention model over feed-forward baselines and other state-of-the-art methods for monocular depth estimation on two challenging datasets – KITTI and DIODE. We show that our proposed approach either outperforms or performs at least on a par with the state-of-the-art monocular depth estimation methods with less memory and computational complexity.**

I. Introduction

Depth perception in human brain requires combined signals from three different cues (oculomotor, monocular, and binocular [1]) to obtain ample information for depth sensing and navigation. As a matter of fact, the domain of MDE is inherently ill-posed. Classical computer vision approaches employ multi-view stereo correspondence algorithms [2], [3] for depth estimation. Recent deep learning based methods formulate MDE as a dense, pixel-level, continuous regression problem [4], [5], although a few approaches pose it as a classification [6] or quantized regression [7] task.

State-of-the-art MDE models [4], [5], [7] are based on pretrained convolutional backbones [8]–[10] with upsampling and skip connections [4], global context module and logarithmic discretization for ordinal regression [7], and coefficient learner for upsampling with local planar assumption [5]. All these design choices explicitly or implicitly suffer from the spatial downsampling operation in the backbone architecture which is shown for pixel-level task [11]. We incorporate the idea of depth-to-space (D2S) [12], [13] transformation as a remedy to the downsampling operation in the decoding phase. However, straightforward D2S transformation of the final feature map might suffer from the lack of global context from the scene necessary for reliable estimation [14], [15]. Therefore, we inject global context on the D2S transformed single-channel feature maps for each stage of the backbone.

Moreover, we effectively gather information from all the stages of the backbone with a bidirectional attention module (see Fig. 2) inspired by [16] which first demonstrates the potential of attention in deep learning. Recently proposed attention-based works [17]–[21] mostly exploit either channel-wise or spatial attention in some form.

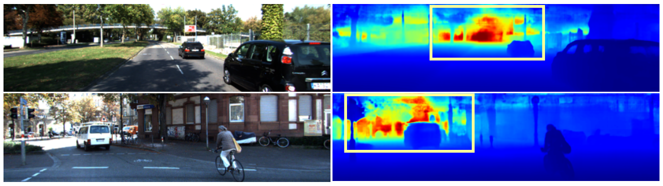

Fig. 1: Sample BANet predictions on KITTI *val* set. BANet improves the overall depth estimation by generating attention weights for each stage and diminishing representational ambiguity within the network.

These approaches manipulate the internal feature maps of the backbone architecture with higher depth dimensions. For a RGB image with resolution, $H \times W \times 3$, the attention mechanism is applied on $\frac{H}{s} \times \frac{W}{s} \times D$ with stride, $s \in \{2, 4, 8, 16, 32\}$ and much higher depth D (generally between 256 and 2208). We hypothesize that the key aspect of a more accurate MDE architecture is its ability to produce the output with the same spatial resolution as input. Therefore, in our network design, we first transform all the backbone feature maps of dimension, $\frac{H}{s} \times \frac{W}{s} \times D$ into the original resolution of $H \times W \times 1$ followed by refinement process using our proposed bidirectional attention mechanism. Unlike other attention based approaches [17]–[21], our bidirectional refinement process is applied outside the backbone with the original spatial resolution. This is efficiently doable for MDE as it is a single plane estimation task unlike the multi-class segmentation problem. Moreover, running the bidirectional attention modules on top of the single channel feature map brings additional advantages including computational efficiency and lower number of parameters.

Thus, in the light of the highlighted issues that arise in MDE methods, we present a novel yet effective pipeline for estimating continuous depth map from a single image (see Fig. 1). Although our architecture contains substantially more connections than SOTA, both the computational complexity and the number of parameters are lower than the recent approaches as most of the interactions are computed on D2S transformed single channel features. Due to the prominence of bidirectional attention pattern in our design, we name our model *Bidirectional Attention Network (BANet)*.

Contributions: Our main contributions are as follows:

- To the best of our knowledge, we present the first work applying the concept of bidirectional attention for monocular depth estimation. The flexibility of our mechanism is its ability to be incorporated with any existing CNNs.
- We further introduce *forward* and *backward* attention

*Equal contribution; †corresponding author

Authors are with Noah's Ark Laboratory, Huawei Technologies, Markham, ON L3R5Y1, Canada.
Correspondence: liu.bingbing@huawei.com

978-1-7281-9077-8/21/$31.00 ©2021 IEEE

modules which effectively integrate local and global information to filter out ambiguity.
- We present an extensive set of experimental results on two of the large scale MDE datasets to date. The experiments demonstrate the effectiveness of our proposed method in both the efficiency as well as performance. We show that variants of our proposed mechanism perform better or on a par with the SOTA architectures.

II. Related Work

Supervised MDE: A compiled survey of classical stereo correspondence algorithms for dense disparity estimation is provided in [2]. Saxena *et al.* [22] employed Markov Random Field (MRF) to extract absolute and relative depth magnitudes using the local cues at multiple scales to exploit both local and global information from image patches. Based upon the assumption that the 3D scenes comprise many small planar surfaces (i.e. triangulation), this work is further extended to estimate the 3D position and orientation of the oversegmented superpixels in the images [23], [24].

Eigen *et al.* [25] is the first work addressing the monocular depth estimation task with deep learning. The authors proposed a stack of global, coarse depth prediction model followed by local region-wise refinement with a separate sub-network. Li *et al.* [26] refined the superpixel depth map predicted by the CNN model to the pixel-level using a CRF. Cao *et al.* [6] transformed the continuous regression task of MDE to the pixel-level classification problem by discretization of depth ranges into multiple bins. A unified approach blending CRF and fully convolutional networks with superpixel pooling for faster inference is proposed into the framework of Deep Convolutional Neural Fields (DCNF) [27]. Multi-scale outputs from different stages of CNN are fused with continuous CRFs in [28] with further refinements via structured attention over the feature maps [29].

Recently, DORN [7] modeled the MDE task as an ordinal regression problem with space-increasing (logarithmic) discretization of the depth range to appropriately address the increase in error with depth magnitude. DenseDepth [4] employed the pretrained DenseNet [10] backbone with bilinear upsampling and skip connections on the decoder to obtain high-resolution depth map. The novelty in the BTS architecture [5] lies in their local planar guidance (LPG) module as an alternative to upsampling based skip connections to transform intermediate feature maps to full-resolution depth predictions. In contrast, we propose a lightweight bidirectional attention mechanism to filter out ambiguity from the deeper representations by incorporating global and local contexts.

Visualization of MDE models: Hu *et al.* [15] showed that CNNs most importantly need the edge pixels to infer the scene geometry and regions surrounding the vanishing points for outdoor scenarios. Dijk and Croon [14] found that the pretrained MDE models ignore the apparent size of the objects and focus only on the vertical position of the object in the scene to predict the distance. It means that the models approximate the camera pose while training. This is particularly true for the KITTI dataset for which the camera is mounted on top of the vehicle and the vanishing point of all the scenes is around the middle of the images. Consequently, a small change in pitch corrupts the depth prediction in a non-negligible manner.

III. Our Approach

In this section, we discuss our proposed Bidirectional Attention Network (BANet). Specifically, we introduce bidirectional attention modules and different global context aggregation techniques for effective integration of local and global information for the task of monocular depth estimation.

A. Bidirectional Attention Network

The idea of applying bidirectional attention in our proposed approach is motivated by neural machine translation (NMT) [16]. Although there exist recent works [17]–[21] that exploits channel-wise and spatial attention in CNN for various computer vision tasks, the idea of applying attention in forward and backward manner to achieve the nature of birectional RNN has not been explored yet.

Our overall architecture is illustrated in Fig. 2. In general, standard CNNs for image classification [8]–[10], [17], [30], [31] consists of five feed-forward stages. We denote the final feature map from stage i as s_i, thus resulting in an ordered set of five feature maps, $\{s_1, s_2, s_3, s_4, s_5\}$. To draw an analogy with the words in a sentence used in NMT [16], this set can be portrayed as a visual sentence (ignoring their input-output dependencies) where each s_i is analogous to a word. Note that the bidirectional RNN inherently generates the forward and backward hidden states due to the sequential and dynamic processing of the source sentence word by word. Since CNNs have a static nature on the input images, we introduce forward and backward attention sub-modules which take the stage-wise feature maps, s_i as input.

Bidirectional Attention Modules: First, we reshape the stage-wise feature maps, s_i, (where $i = 1, 2, .., 5$) to the output spatial resolution using a 1×1 convolution layer followed by the depth-to-space (D2S) [12], [13] operation. Our forward and backward attentions are given by:

$$\begin{aligned}
s_i^f &= \mathcal{D}_i^f\left(s_i\right); & s_i^b &= \mathcal{D}_i^b\left(s_i\right)\\
a_i^f &= \mathcal{A}_i^f\left(s_1^f, s_2^f, \ldots, s_i^f\right); & a_i^b &= \mathcal{A}_i^b\left(s_i^b, s_{i+1}^b, \ldots, s_N^b\right)\\
\mathcal{G}^f &= a_1^f \otimes a_2^f \otimes \ldots \otimes a_N^f; & \mathcal{G}^b &= a_1^b \otimes a_2^b \otimes \ldots \otimes a_N^b\\
\mathcal{A} &= \varphi\left(\mathcal{G}\left(\mathcal{G}^f \otimes \mathcal{G}^b\right)\right) & &
\end{aligned} \tag{1}$$

$\forall i \in \{1, 2, \ldots, N\}$. In Eq. 1, the superscripts $\square^f$ and $\square^b$ denote the operations related to the forward and backward attentions, respectively, and the subscript $\square_i$ denotes the associated stage of the backbone feature maps. $\mathcal{D}_i^f$ and $\mathcal{D}_i^b$ represent the $D2S$ modules which process the backbone feature map s_i. $\mathcal{A}_i^f$ and $\mathcal{A}_i^b$ refer to the 9×9 convolution in Fig. 2. Note that $\mathcal{A}_i^f$ has access to the features up to the i^{th} stage, and $\mathcal{A}_i^b$ receives the feature representation from the i^{th} stage onward; thus emulating the forward and backward

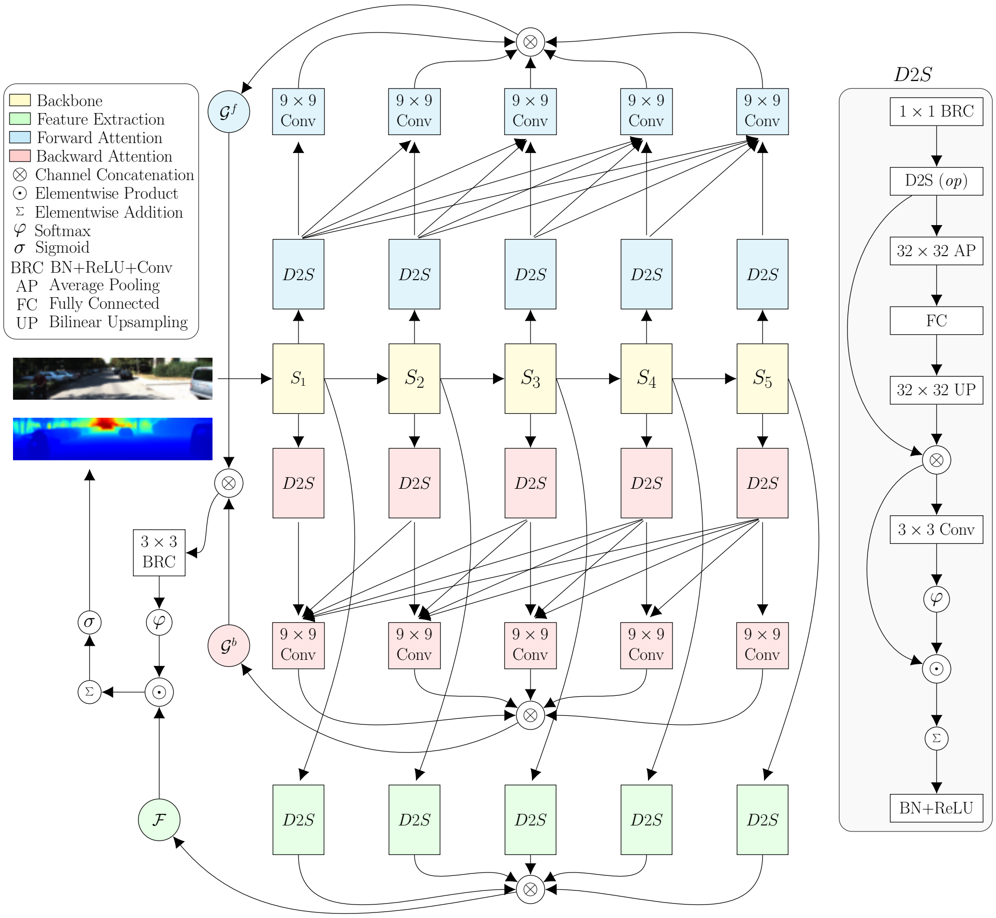

Fig. 2: The computation graph of our proposed bidirectional attention network (BANet). The directed arrows refer to the flow of data through this graph. The forward and backward attention modules take stage-wise features as input and process through multiple operations to generate stage-wise attention weights. The rectangles represent either a single or composite operation with learnable parameters. The white circles with inscribed symbols ($\otimes, \varphi, \odot, \sum, \sigma$) denote parameterless operations. The colored circles highlight the outputs of different conceptual stages of processing. The construction of the $D2S$ modules is shown inside the gray box on the right. Note that the green $D2S$ modules used for feature computation $\mathcal{F}$ does not employ the last BN+ReLU block inside $D2S$.

attention mechanisms of a bidirectional RNN. Next, all the forward and backward attention maps are concatenated channel-wise and processed with a 3×3 convolution ($\mathcal{G}$ in Eq. 1) and softmax (φ) to generate the per stage pixel-level attention weights, $\mathcal{A}$ (see Fig. 3). The procedure of computing the feature representation, f_i, from the stage-wise feature maps, S_i, using the D2S module can be formalized as:

$$f_i = \mathcal{D}_i^{\mathcal{F}}(s_i); \quad \mathcal{F} = f_1 \otimes f_2 \otimes \ldots \otimes f_N \qquad (2)$$

$\forall i \in \{1, 2, \ldots, N\}$. Then we compute the unnormalized prediction, $\widehat{D}_u$, from the concatenated features, $\mathcal{F}$, and attention maps, $\mathcal{A}$, using the Hadamard (element-wise) product followed by pixel-wise summation $\sum$. Finally, the normalized prediction, $\widehat{D}$, is generated using the sigmoid (σ) function. The operations can be expressed as:

$$\widehat{D}_u = \sum \mathcal{A} \odot \mathcal{F}, \quad \widehat{D} = \sigma\left(\widehat{D}_u\right) \qquad (3)$$

Global Context Aggregation. We incorporate global context inside the $D2S$ modules by applying average pooling with relatively large kernels followed by fully connected layers and bilinear upsampling operations. This aggregation of global context in the D2S module helps to resolve ambiguities (see Fig. 4 and 5) for thinner objects in more challenging scenarios (i.e., very bright or dark contexts). Additionally, we provide details of few alternative realizations of our proposed architecture as follows:

- **BANet-Full:** This is the complete realization of the architecture as shown in Fig. 2.
- **BANet-Vanilla:** It contains only the backbone followed by 1×1 convolution, a single D2S operation, and sigmoid to generate the final depth prediction. This is similar to [32].
- **BANet-Forward**: The backward attention modules of BANet are missing in this setup.
- **BANet-Backward:** The forward attention modules of BANet are not included here.
- **BANet-Markov:** This follows the Markovian assumption that the feature at each time step (or stage) i depends only on the feature from the immediate preceding (for forward

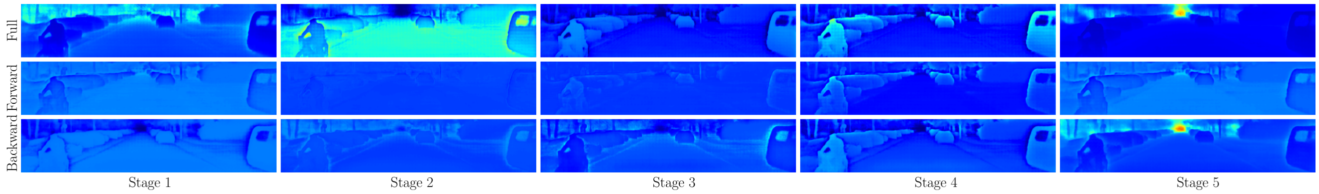

Fig. 3: An illustration of the generated stage-wise attention weights using the forward and backward attention modules.

attention) or succeeding (for backward attention) time step (or stage) $i \mp 1$ (i.e. $X_i \perp \{X_{i\mp2}, X_{i\mp3}, ...X_{i\mp k}\}|X_{i\mp1}$).

- **BANet-Local:** This variant replaces the global context aggregation part with a single 9×9 convolution.

To demonstrate that the performance improvement on the MDE task using the time dependent structuring is not due to the mere increase in the capacity as well as the number of parameters, we conduct pilot experiments without any time dependent structuring by simply concatenating the different stage features all at once followed by similar post-processing. However, we empirically find that such a naive implementation performs much poorer than our proposed time dependent realizations mentioned above. Therefore, we exclude this straightforward employment from further experimental analysis.

IV. Experiments

Datasets: Among several publicly available standard datasets [23], [24], [33]–[35] for monocular depth estimation, we choose DIODE [34] and KITTI [35] due to their comparatively high density of ground truth annotations. We hypothesize that our model assessment might be biased towards better nonlinear interpolator rather than the candidates capable of learning to predict depth maps using monocular cues. Therefore, our experiments done on the datasets comprising truly dense ground truths provide unbiased measure of different types of architectural design choices.

- **DIODE**: Dense Indoor/Outdoor DEpth (DIODE) [34] is the first standard dataset for monocular depth estimation comprising diverse indoor and outdoor scenes acquired with the same hardware setup. The training set consists of 8574 indoor and 16884 outdoor samples from 20 scans each. The validation set contains 325 indoor and 446 outdoor samples with each set from 10 different scans. The ground truth density for the indoor training and validation splits are approximately 99.54% and 99%, respectively. The density of the outdoor sets are naturally lower with 67.19% for training and 78.33% for validation subsets. The indoor and outdoor ranges for the dataset are $50m$ and $300m$, respectively.
- **KITTI**: KITTI dataset for monocular depth estimation [35] is a derivative of the KITTI sequences. The training, validation, and test sets comprise 85898, 1000, and 500 samples, respectively. Following previous works, we set the prediction range to $[0-80m]$ where ever applicable, although there are a few pixels with depth values greater than $80m$.

Implementation Details: For fair comparison, all the recent models [4], [5], [7] are trained from scratch alongside ours using the hyperparameter settings proposed in the corresponding papers. Our models are trained for 50 epochs with batch size of 32 in the `NVIDIA Tesla V100 32GB GPU` using the sparse labels without any densification. The runtime was measured with a single `NVIDIA GeForce RTX 2080 Ti GPU` just because of its availability at the time of writing. We use Adam optimizer with the initial learning rate of $1e^{-4}$ and weight decay of 0. The learning rate is multiplied on plateau after 10 epochs by 0.1 until $1e^{-5}$. Random horizontal flipping and color jitter are the only augmentations used in training. For KITTI, we randomly crop 352×1216 training samples following the resolution of the validation and test splits. Also, for computational efficiency, the inputs are half-sampled (and zero-padded if necessary) prior to feeding into the models, and the predictions are bilinearly upsampled to match the ground truth before comparison. For DORN [7], we set the number of ordinal levels to the optimal value of 80 (or 160 considering both positive and negative planes). This value is set to 50, 75, and 75 for DIODE indoor, outdoor, and combined splits following the same principle. We employ SILog and $\mathcal{L}_1$ losses for training all the models on KITTI and DIODE, respectively, except for DORN that comes with its own ordinal loss. Following BTS [5] and DenseDepth [4], we use DenseNet161 [10] as our BANet backbone.

Evaluation Metrics: For the error (lower is better) metrics, we mostly follow the measures (SILog, SqRel, AbsRel, MAE, RMSE, iRMSE) directly from the KITTI leaderboard. Additionally, we have used the following thresholded accuracy (higher is better) metrics (Eq. 4).

$$\delta_k = \frac{1}{N}\sum_i \left(max(\frac{a_i}{t_i}\frac{t_i}{a_i}) < (1 + \frac{k}{100}) \right) * 100 \;\; (\%); \\ k \in \{25, 56, 95\} \tag{4}$$

A. Results on Monocular Depth Estimation

Quantitative Comparison: Table I, II, III, and IV present the comparison of our BANet variants with SOTA architectures. Note that DORN performs much worse due to the effect of granularity provided by its discretization or ordinal levels. A higher value of this parameter might improve its precision, but with exponential increase in memory consumption during training; thus making it difficult to train on large datasets like KITTI and DIODE. Overall, our BANet variants perform better or close to existing SOTA in particular BTS. Finally, our model performs better than SOTA architectures on the KITTI leaderboard [1] (Table V) in terms of the primary metric (SILog) used for ranking.

Qualitative Comparison: Figure 4 and 5 show the qualitative comparison of different SOTA methods. Also, we

[1] http://www.cvlibs.net/datasets/kitti/eval_depth.php?benchmark=depth_prediction

TABLE I: Quantitative results on the **DIODE Indoor** *val* set.

Model	#Params	Time	Lower is better						Higher is better (%)		
	(10^6)	(ms)	**SiLog**	**SqRel**	**AbsRel**	**MAE**	**RMSE**	**iRMSE**	δ_{25}	δ_{56}	δ_{95}
DORN [7]	91.21	57	24.45	49.54	47.25	1.63	1.86	208.30	37.61	62.39	77.97
DenseDepth [4]	44.61	27	19.67	25.10	34.50	**1.39**	**1.56**	177.33	48.13	70.19	83.00
BTS [5]	47.00	34	19.19	25.55	33.31	1.42	1.61	**171.41**	50.14	**72.71**	**83.44**
BANet-Vanilla	28.73	21	18.78	27.75	34.36	1.43	1.61	175.35	**51.28**	70.63	81.32
BANet-Forward	32.58	28	18.39	25.81	35.49	1.49	1.65	179.58	47.53	68.63	80.91
BANet-Backward	32.58	28	**18.25**	24.43	34.19	1.46	1.62	176.02	48.94	69.50	81.52
BANet-Markov	35.64	31	18.62	**21.78**	**32.58**	1.44	1.61	173.58	48.99	72.28	82.29
BANet-Local	35.08	32	18.83	23.65	33.49	1.43	1.59	172.11	49.71	71.18	82.88
BANet-Full	35.64	33	18.43	26.49	34.99	1.45	1.61	177.22	48.48	70.66	82.26

TABLE II: Quantitative results on the **DIODE Outdoor** *val* set.

Model	#Params	Time	Lower is better						Higher is better (%)		
	(10^6)	(ms)	**SiLog**	**SqRel**	**AbsRel**	**MAE**	**RMSE**	**iRMSE**	δ_{25}	δ_{56}	δ_{95}
DORN [7]	91.31	57	40.96	249.72	48.42	5.17	8.06	98.01	46.92	75.36	87.30
DenseDepth [4]	44.61	27	37.49	244.47	41.74	4.32	7.03	110.19	58.56	81.14	90.05
BTS [5]	47.00	34	39.91	236.51	41.86	4.40	7.15	93.00	57.71	80.62	89.73
BANet-Vanilla	28.73	21	37.38	217.52	39.53	4.19	6.93	92.85	58.95	81.42	90.09
BANet-Forward	32.58	28	37.59	**202.70**	38.80	4.25	7.00	93.40	58.46	80.81	89.72
BANet-Backward	32.58	28	36.89	215.27	39.15	4.21	6.86	92.44	57.89	81.34	90.37
BANet-Markov	35.64	31	37.61	226.07	39.86	4.27	6.97	92.86	58.46	81.24	90.00
BANet-Local	35.08	32	**36.54**	218.54	39.23	**4.11**	**6.78**	92.72	**59.68**	**82.35**	**90.56**
BANet-Full	35.64	33	37.17	210.97	**38.50**	4.20	6.94	**92.26**	58.66	81.53	90.33

TABLE III: Quantitative results on the **DIODE All** (Indoor+Outdoor) *val* set.

Model	#Params	Time	Lower is better						Higher is better (%)		
	(10^6)	(ms)	**SiLog**	**SqRel**	**AbsRel**	**MAE**	**RMSE**	**iRMSE**	δ_{25}	δ_{56}	δ_{95}
DORN [7]	91.31	57	34.23	226.87	47.30	3.55	5.34	135.86	46.41	73.54	86.02
DenseDepth [4]	44.61	27	31.05	157.70	39.33	3.10	4.77	182.26	54.25	76.46	87.24
BTS [5]	47.00	34	29.83	**146.12**	**37.03**	3.01	4.71	152.42	**56.35**	77.29	**88.14**
BANet-Vanilla	28.73	21	30.11	156.16	39.30	3.08	4.73	129.02	53.76	76.67	87.37
BANet-Forward	32.58	28	30.54	181.53	39.09	3.05	4.77	128.88	55.17	76.86	87.21
BANet-Backward	32.58	28	30.28	179.41	40.25	3.05	4.73	129.98	54.17	76.83	87.47
BANet-Markov	35.64	31	30.16	188.45	40.10	3.12	4.80	129.31	54.04	76.04	86.76
BANet-Local	35.08	32	**29.69**	148.91	37.68	**2.98**	**4.66**	**126.96**	56.08	**77.34**	87.96
BANet-Full	35.64	33	29.93	150.76	39.28	3.06	4.73	130.39	53.26	76.40	87.15

TABLE IV: Quantitative results on the **KITTI** *val* set.

Model	#Params	Time	Lower is better						Higher is better (%)		
	(10^6)	(ms)	**SiLog**	**SqRel**	**AbsRel**	**MAE**	**RMSE**	**iRMSE**	δ_{25}	δ_{56}	δ_{95}
DORN [7]	89.23	36	12.32	2.85	11.55	1.99	3.69	11.71	91.60	98.47	99.55
DenseDepth [4]	44.61	25	10.66	1.76	8.01	**1.58**	3.31	8.24	93.50	**98.87**	99.69
BTS [5]	47.00	29	10.67	1.63	**7.64**	**1.58**	3.36	**8.14**	93.63	98.79	99.67
BANet-Vanilla	28.73	20	10.88	1.65	7.74	1.63	3.51	8.34	93.49	98.63	99.59
BANet-Forward	32.34	25	10.61	1.67	9.02	1.77	3.54	9.99	92.66	98.68	99.61
BANet-Backward	32.34	25	10.54	**1.52**	7.67	1.59	3.42	8.47	93.57	98.72	99.65
BANet-Markov	35.28	27	10.72	1.85	8.24	1.62	3.38	8.28	93.19	98.75	99.69
BANet-Local	35.08	27	**10.53**	2.21	9.92	1.77	3.34	9.68	92.36	**98.87**	**99.71**
BANet-Full	35.28	28	10.64	1.81	8.25	1.60	**3.30**	8.47	**93.75**	98.81	99.68

TABLE V: Results on the KITTI leaderboard (test set).

Method	SILog	SqRel	AbsRel	iRMSE
DORN [7]	11.77	2.23	**8.78**	12.98
BTS [5]	11.67	**2.21**	9.04	12.23
BANet-Full	**11.55**	2.31	9.34	**12.17**

provide a video of our predictions for the KITTI validation sequences [2]. The DORN prediction map clearly indicates its semi-discretized nature of prediction. For both the indoor and outdoor images in Fig. 4, BTS and DenseDepth suffer from the high intensity regions surrounding the window frame in the bottom right of the left image and the trunks in the right image. Our BANet variants work better in these delicate regions. To better tease out the effect of the different components in BANet, we provide observations as follows:

- **Why does vanilla perform better than SOTA in many cases?** The vanilla architecture inherently maintains and generates the full-resolution depth map using the D2S conversion without increasing the number of parameters. We hypothesize this to be the primary reason why it performs better than DORN (huge parameters) and DenseDepth (generates $4\times$ downsampled output followed by $4\times$ upsampling).
- **Why is Markovian worse than vanilla?** BANet-Markov utilizes only the S_{i-1} stage features to generate the attention map for S_i stage features. For other variants, when $s_{i-1}, s_{i-2}, s_{i-3}, \ldots$ are used for the attention mechanism, the forgotten (or future for backward) information from $s_{i-2}, s_{i-3}, \ldots$ are used for the refinement. However, essentially no forgotten information is brought back to the stage for Markovian formulation. The same information (or s_{i-1})

[2] https://www.youtube.com/watch?v=a9wW02pfeKw&ab_channel=MannatKaur

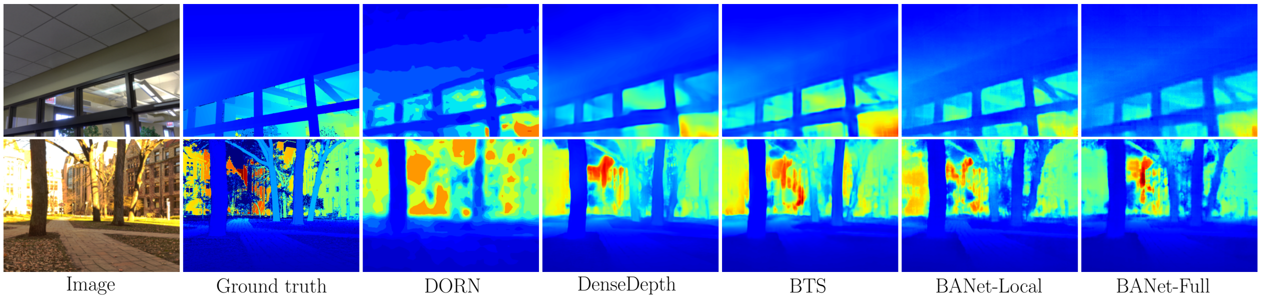

Fig. 4: Sample results on DIODE indoor (top) and outdoor (bottom) images. **Top:** The window frame on the bottom right is detected well by BANets compared to BTS and DenseDepth. **Bottom:** Other methods but our BANet-Full fail to localize the highly illuminated tree trunks. This shows the importance of global context aggregation in BANet-Full.

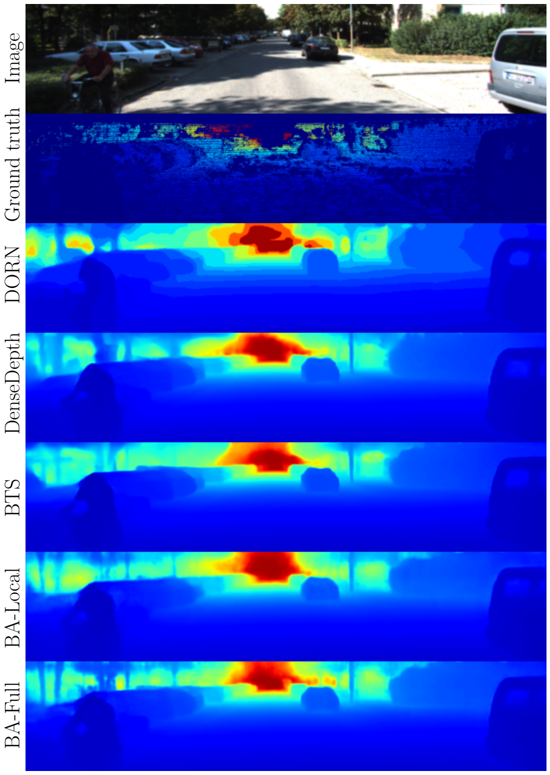

Fig. 5: Sample results on KITTI *val* set. The top region is cropped since it does not contain any ground truth. The tree trunks in the dark regions on the left is better differentiated with global context aggregation in BANet-Full.

that was directly used to compute the feature map s_i is also being reused for attention. We believe this lack of forgotten information prevent the BANet-Markov variant to gain a clear edge compared to the vanilla architecture.

• **Backward refinement is more crucial than the forward:** The superior performance of BANet-Backward compared to BANet-Forward indicates that the feature refinement is potentially more important in the early stages of the backbone than later in terms of performance.

• **Effect of global context aggregation:** The effect of global context aggregation cannot be well-understood with the numerical differences in performance. This is because it mostly attempts to refine the prediction for the delicate regions like the thin tree trunks and branches both in extremely dark or bright backgrounds (Figure 4 and 5). Note that the pixel coverage of such critical regions is quite low compared to the total number of annotated pixels, and the vanilla and local variants perform well for these (comparatively easier) regions. However, in a real scenario, it is more important to identify the related depth of both the difficult and easy pixels satisfactorily than to be better only at the easy pixels, which are in bulk quantities in all the datasets. In other words, compared to the other variants, BANet-Full has a slightly higher residual error with much less number of false negatives which is critical for sensitive applications like autonomous driving. Therefore, based upon our understanding, BANet-Full is a more reliable alternative than the others in almost all the cases even though it is slightly off in numbers.

• **General pattern of stage-wise attention:** The backward attention dominates the forward ones (Figure 3) because the raw, stage-wise feature maps are computed with the backbone in the forward direction, thus encoding more forward information. The general pattern of stage-wise bidirectional attention can be described as follows:

Stage-1: Generic activations spread over the whole scene.

Stage-2: Emphasis is on the ground plane or road regions with gradual change in depth from the observer.

Stage-3: Focus is on the individual objects, such as cars, pedestrians, etc. This one assists in measuring the ground plane distance of the objects from the observer's location.

Stage-4: Focus is still on the objects with a gradual change within each object from the ground towards the top. Therefore, this is measuring the intra-object depth variation.

Stage-5: Primarily devoted to the understanding of the vanishing point information to get a sense of the upper bound of depth in the scene in general.

V. Conclusion

In this paper, we present a bidirectional attention mechanism for convolutional networks for monocular depth estimation task similar to the bidirectional RNN for NMT. We show that a standard backbone equipped with this attention mechanism performs better or at least on a par with the SOTA architectures on the standard datasets. As a future direction, we believe that our proposed attention mechanism has the potential to be useful in similar dense prediction tasks, esp. the single plane ones.

References

[1] E. Goldstein, *Sensation and Perception*. Cengage Learning, 2009. 1

[2] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *IJCV*, vol. 47, 2002. 1, 2

[3] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *CVPR*, 2006. 1

[4] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *CoRR*, vol. abs/1812.11941, 2018. 1, 2, 4, 5

[5] J. H. Lee, M. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," *CoRR*, vol. abs/1907.10326, 2019. 1, 2, 4, 5

[6] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, 2018. 1, 2

[7] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *CVPR*, 2018. 1, 2, 4, 5

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015. 1, 2

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016. 1, 2

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017. 1, 2, 4

[11] M. Amirul Islam, M. Rochan, N. D. Bruce, and Y. Wang, "Gated feedback refinement network for dense image labeling," in *CVPR*, 2017. 1

[12] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *CVPR*, 2016. 1, 2

[13] S. Aich, W. van der Kamp, and I. Stavness, "Semantic binary segmentation using convolutional networks without decoders," in *CVPRW*, 2018. 1, 2

[14] T. v. Dijk and G. d. Croon, "How do neural networks see depth in single images?" in *ICCV*, 2019. 1, 2

[15] J. Hu, Y. Zhang, and T. Okatani, "Visualization of convolutional neural networks for monocular depth estimation," in *ICCV*, 2019. 1, 2

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015. 1, 2

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018. 1, 2

[18] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018. 1, 2

[19] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *CVPR*, 2018. 1, 2

[20] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *CVPR*, 2020. 1, 2

[21] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *ICCV*, 2019. 1, 2

[22] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *NIPS*, 2006. 2

[23] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *NIPS*, 2006. 2, 4

[24] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 824–840, 2009. 2, 4

[25] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *NIPS*, 2014. 2

[26] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs," in *CVPR*, 2015. 2

[27] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE TPAMI*, vol. 38, no. 10, 2016. 2

[28] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," in *CVPR*, 2017. 2

[29] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *CVPR*, 2018. 2

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012. 2

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016. 2

[32] J. M. U. Vianney, S. Aich, and B. Liu, "Refinedmpl: Refined monocular pseudolidar for 3d object detection in autonomous driving," 2019. 3

[33] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012. 4

[34] I. Vasiljevic, N. Kolkin, S. Zhang, R. Luo, H. Wang, F. Z. Dai, A. F. Daniele, M. Mostajabi, S. Basart, M. R. Walter, and G. Shakhnarovich, "DIODE: A Dense Indoor and Outdoor DEpth Dataset," *CoRR*, vol. abs/1908.00463, 2019. 4

[35] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," in *3DV*, 2017. 4