

DEPTHFORMER: MULTISCALE VISION TRANSFORMER FOR MONOCULAR DEPTH ESTIMATION WITH GLOBAL LOCAL INFORMATION FUSION

Ashutosh Agarwal Chetan Arora

Indian Institute of Technology Delhi

ABSTRACT

Attention-based models such as transformers have shown outstanding performance on dense prediction tasks, such as semantic segmentation, owing to their capability of capturing long-range dependency in an image. However, the benefit of transformers for monocular depth prediction has seldom been explored so far. This paper benchmarks various transformer-based models for the depth estimation task on an indoor NYUV2 dataset and an outdoor KITTI dataset. We propose a novel attention-based architecture, *Depthformer* for monocular depth estimation that uses multi-head self-attention to produce the multiscale feature maps, which are effectively combined by our proposed decoder network. We also propose a *Transbins* module that divides the depth range into bins whose center value is estimated adaptively per image. The final depth estimated is a linear combination of bin centers for each pixel. *Transbins* module takes advantage of the global receptive field using the transformer module in the encoding stage. Experimental results on NYUV2 and KITTI depth estimation benchmark demonstrate that our proposed method improves the state-of-the-art by **3.3%**, and **3.3%** respectively in terms of Root Mean Squared Error (RMSE). Code is available at <https://github.com/ashutosh1807/Depthformer.git>.

Index Terms— depth estimation, transformer, attention, adaptive bins

1. INTRODUCTION

Depth estimation from a single image is fundamental for many applications, from virtual reality to low-cost autonomous driving. Almost all of the current techniques to estimate depth from a single image are based on convolutional neural networks (CNN) with a U-Net-based encoder-decoder architecture [1, 2, 3, 4, 5, 6, 7, 8, 9]. The encoder is typically an image classification network trained on Imagenet [10], and the decoder aggregates multiscale features to produce final dense depth. As earlier works have pointed out, the features extracted from a CNN have a local receptive field [11, 12]. For dense prediction tasks such as semantic segmentation and depth estimation, a pixel must have a global receptive field about the scene along with the local information for a more

accurate estimation.

The recent success of transformers in Natural Language Processing tasks has created considerable interest among researchers to introduce them in computer vision tasks owing to their capability of capturing long-range dependencies. However, earlier works in Transformer based vision architectures mainly focused on Classification and Object Detection [13, 14]. With the advent of Multiscale Vision Transformers (MVTs), researchers have also started using transformer-based architectures as encoders for dense prediction tasks like semantic segmentation [12, 15, 11, 16, 17]. Most of the works using MVTs showcase their dense prediction performance for semantic segmentation. On the contrary, depth estimation is even more difficult for two reasons: (1) It is a continuous prediction task, and (2) It is an ill-posed problem owing to the inherent scale ambiguity. To the best of our knowledge [18] is the only work that proposes to use vision transformer (ViT) during the encoding stage. However, it applies ViT only on feature maps with $1/16$ th scale extracted using a CNN architecture. Motivated by these observations, we propose a novel transformer-based encoder for depth estimation that uses multiheaded self-attention to produce hierarchical feature maps. We also benchmark the performance of recently proposed vision transformers (ViTs) for semantic segmentation on the monocular depth estimation task.

Using a transformer-based encoder increases the receptive field of the network, but for the dense prediction task, pixels also must understand local information. For example, a pixel must know that it lies on a boundary of an object or that it belongs to a group of pixels on a co-planar surface. Previous works that use MVTs [11] as encoders have a decoder design in which they upscale the encoded features of varying resolutions to a fixed resolution and fuse them using MLP layers. However, directly upsampling to a higher resolution and fusion results in a loss in local information. Motivated by this, we propose a novel decoder that iteratively upsamples feature maps and fuses them with the encoder features, starting from the lowest resolution and moving towards the high resolution. Iterative upsampling and fusion helps propagate global information to high-resolution local information preserving features.

In alignment with the state-of-the-art (SoTA) [19], we model the depth estimation task as an ordinal regression

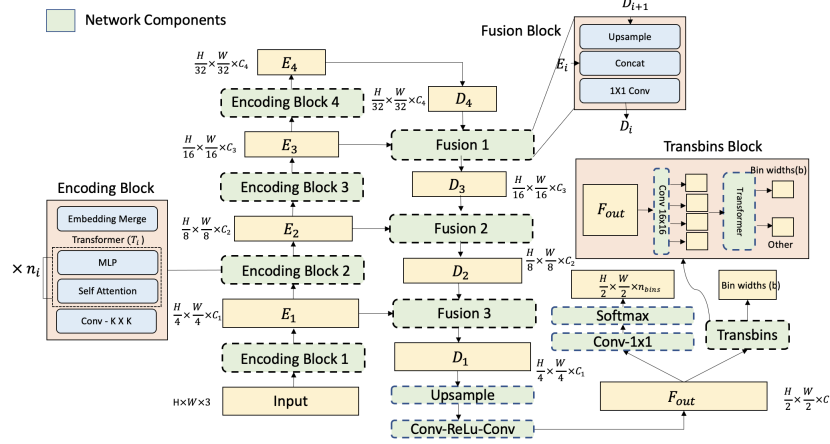


Fig. 1. The architecture of our proposed method, Depthformer : MVT produces feature maps of multiple resolutions hierarchically fused by the decoder network to produce output feature map F_{out} . F_{out} is fed to the Transbins module predicts the bin widths. Pixel-wise probability distribution over the bins centers is finally predicted by using a 1×1 convolution followed by *softmax* activation.

problem that discretizes the depth range into a number of intervals or bins adaptively per image. We propose a novel module Transbins, that takes advantage of fused global information into the feature maps by an MVT and produces adaptive bins discretized over the depth range. To demonstrate the efficacy of our proposed network Depthformer, we conduct extensive experiments on two canonical depth estimation benchmarks, an outdoor driving dataset KITTI [20] and an indoor dataset NYUV2 [21], achieving SoTA on both.

Contributions: To summarise, our contributions are - (1) We propose a novel encoder-decoder network that uses self-attention to predict multiscale feature maps that are effectively fused by our decoder. (2) We propose a novel Transbins module that predicts adaptive bins discretized over the depth range whose centers are used to predict final depth. (3) We benchmark the performance of earlier proposed multiscale vision transformers (MVTs) on monocular depth estimation. (4) Our network Depthformer achieves state-of-the-art performance on both outdoor dataset KITTI and indoor dataset NYUV2.

2. METHODOLOGY

Problem Definition: Following [19, 2], we model depth estimation as an ordinal regression task. Given an input I , our network predicts: (1) Bin widths, b , that discretize continuous depth range into a number of intervals, n_{bins} , adaptively per image. (2) Pixel-wise probability distribution over the adaptive bins. The final depth, d , at a pixel is the linear combination of the probability scores at the pixel and per image depth-bin-centers.

Architecture: This section introduces the architecture for the encoder and decoder of our proposed Depthformer and the

training loss that we follow. Our encoder produces the feature maps of varying resolutions using a MVT given an input image. The decoder then fuses these multiscale features to predict feature map F_{out} . Next, the proposed Transbins module uses F_{out} as input to predict adaptive bins b discretized over the depth range. The probability distribution over the bins is estimated by applying a convolution operation on F_{out} followed by the *softmax* activation. We describe a more detailed architecture below.

Encoder: Given an image I of size $H \times W \times 3$, our encoder aims to create feature maps at multiple resolutions using the Transformer framework. Input I is first convolved with learned C_1 kernels of size K_1 and a stride of 4 to produce a feature map of size $\frac{H}{4} \times \frac{W}{4} \times C_1$. This feature map is flattened to produce a sequence of $\frac{HW}{4^2}$ feature vectors, each of dimension C_1 that are fed to the transformer block T_1 . Transformer T_1 applies n_1 self-attention - MLP layers to produce the output, which is reshaped to a feature map E_1 of size $\frac{H}{4} \times \frac{W}{4} \times C_1$. This process is performed repeatedly, with a stride of 2 for the convolution operation in the following layers. Hence, the encoder produces feature maps E_1, E_2, E_3, E_4 of resolutions $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ respectively with channels $[C_1, C_2, C_3, C_4] = \{64, 128, 320, 512\}$ respectively, that are then fed to the decoder.

Note that, the self attention in the Transformer Block (T_i) is optimised by Spatial attention Reduction (SRA) as suggested in [12]. SRA is implemented via a convolution layer with kernel size R_i and stride R_i to project the key-value pairs, hence resulting in $\frac{n}{R^2}$ compressed key-value pairs where n is the number of input vectors. For the convolution operation with kernel size K , stride S and padding P , we use $K_1 = 7, S_1 = 4, P_1 = 3$ for the first layer and $K_i = 3, S_i = 2, P_i = 1$ for $i = 2, 3, 4$. A kernel size K greater than the stride S , encourages shared information be-

tween the adjacent feature vectors to produce a smoother feature maps. The number of self-attention MLP layers for the four transformer blocks are: $[N_1, N_2, N_3, N_4] = \{3, 8, 27, 3\}$. Spatial reduction ratios for the four transformers are: $[R_1, R_2, R_3, R_4] = \{8, 4, 2, 1\}$. Fig. 1 gives an overview of our encoder design.

Decoder: Earlier works using MVT for dense prediction [11] upsample feature maps at varying resolutions to a resolution of $\frac{1}{4}$, and reduce their channel dimensions to C using 1×1 convolutions. The features are then concatenated and finally fused to predict the output of size $\frac{H}{4} \times \frac{W}{4} \times n_{cls}$ for the segmentation task, where n_{cls} are the number of classes. The feature maps are then upsampled using interpolation to $H \times W \times n_{cls}$ which helps in producing a smoother estimation. Such a decoder design suffers from the loss of local information due to the smoothing effect of interpolation. Earlier works using CNN have used Feature Pyramid Network (FPN) [22] architecture design to preserve the local details. We adopt a similar design and a decoder that iteratively fuses feature maps from the lowest resolution for MVTs.

Effectively, for encoder feature maps E_1, E_2, E_3 , and E_4 with resolution $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$, we iteratively perform the following operation

$$D_i = \text{Conv}\{\text{Concat}[\text{Upsample}(D_{i+1}), E_i]\} \\ i = 1, 2, 3, 4 \quad (1)$$

This procedure produces a map feature map F_{out} of size $\frac{H}{2} \times \frac{W}{2} \times C$ which is fed to the Transbins module as shown in Fig. 1. We select $C = 128$ as in [19]. To upsample the feature maps we have used transposed convolution with kernel size $k = 2$ and stride $s = 2$.

Transbins: Adabins [19] predicts adaptive bins and attentual maps, and fuse the later with the feature map from decoder F_{out} . The motivation is to fuse global information in the attentual maps with the decoder features. We take advantage of the encoded global information in F_{out} via our encoder to only predict bin widths from the full-scale VIT. To predict the distribution over the bins, we use 1×1 convolution over F_{out} followed by *softmax* to predict the output of size $\frac{H}{2} \times \frac{W}{2} \times n_{\text{bins}}$.

The final depth is predicted by linear combination over the bin centers as in [19], which is then upsampled using bilinear interpolation to predict the depth at full resolution. The effectiveness of our proposed Transbins approach against Adabins for bin widths prediction can be seen in Table 4 which reduces the RMSE error by 14.2 %.

Training loss: We train our network on a sum of scaled version of the Scale-Invariant (SI) loss introduced by Eigen *et al.* [1] L_{SILog} and Chamfer loss L_{Chamfer} [19]. L_{SILog} reduces the difference between the predicted depth map and the ground truth depth map. L_{Chamfer} encourages the bin centers to be close to the actual ground truth depth values and vice versa.

$$L_{\text{total}} = L_{\text{SILog}} + \gamma L_{\text{Chamfer}}. \quad (2)$$

Method	RMSE ↓	Rel ↓	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
Eigen <i>et al.</i> [1]	0.641	0.158	0.769	0.95	0.988
DORN [2]	0.509	0.115	0.828	0.965	0.992
Chen <i>et al.</i> [5]	0.514	0.111	0.878	0.977	0.994
VNL [23]	0.416	0.108	0.875	0.976	0.994
BTS [3]	0.392	0.110	0.885	0.978	0.994
DAV [4]	0.412	0.108	0.882	0.980	0.996
DPT-Hybrid [18]	<u>0.357</u>	0.110	0.904	0.988	0.998
Adabins [19]	0.364	<u>0.103</u>	<u>0.903</u>	<u>0.984</u>	<u>0.997</u>
Depthformer (ours)	0.345	0.100	0.911	0.988	<u>0.997</u>

Table 1. Results on NYUV2 Dataset. The best results are in **bold** and second best are underlined. Our method outperforms the previous SoTA methods in most of the metrics.

We have used $\gamma = 0.1$ as in [19].

3. EXPERIMENTS

Datasets: NYU Depth v2 [21] contains indoor scenes dataset with 640×480 resolution images and depth maps with the upper bound of 10 m. We have trained our network on a 24K subset of dataset with a random crop of 576×448 . We evaluate on the pre-defined center cropping by Eigen *et al.* [1] on test set of 654 images.

KITTI [20] is an outdoor scenes dataset containing stereo images with 1241×376 resolution and corresponding 3D laser scans of outdoor scenes. We train our network on a subset of around 26K images from the left view with a random crop of size 704×352 and test on 697 images. To compare our performance, we evaluate on a predefined crop by Garg *et al.* [24] on 697 test images with a maximum value of 80 m.

Evaluation Metrics: We use the standard five metrics used in earlier works [1] to compare our method against SoTA. Given the predicted depth d_p , and the ground truth depth d_p^* at a pixel p , and n denoting the total number of pixels in an image, the error metrics are defined as:

Root mean squared error (RMSE): $\sqrt{\frac{1}{n} \sum_{p=1}^n (d_p - d_p^*)^2}$

Average relative error (REL): $\frac{1}{n} \sum_{p=1}^n \frac{|d_p - d_p^*|}{d_p^*}$

Threshold accuracy (δ_i): % of d_p such that $\max\left(\frac{d_p}{d_p^*}, \frac{d_p^*}{d_p}\right) = \delta_i < \text{Thr}$, where $\text{Thr} = 1.25, 1.25^2, 1.25^3$.

Implementation Details: To train our network, we use the AdamW optimizer [25] with weight-decay 0.1. We follow the training methodology prescribed in [26], and use 1-cycle policy for the learning rate with $\text{max_lr} = 10^{-4}$, linear warm-up from $(3/10)\text{max_lr}$ to max_lr for the first 50% of iterations, followed by cosine annealing to $(3/10)\text{max_lr}$. The network has been trained on 4 V100 GPUs with a 32GB memory, with a batch size of 16 for both NYUV2 and KITTI.

Comparison with State of the Art: Table 1 and 2 demonstrate the performance of our proposed method, Depthformer, with the previous SoTA methods. We consistently outperform

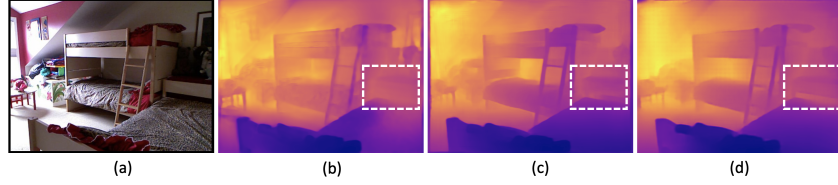


Fig. 2. Visualisation of predicted Depth map for input image (a) for (b) Adabins [19] (c) DPT-Hybrid [18] (d) **Depthformer (Ours)**. Our method is able to predict a more accurate depth estimation owing to its capabilities of capturing long range information.

Method	$RMSE \downarrow$	$Rel \downarrow$	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
Eigen <i>et al.</i> [1]	6.307	0.203	0.702	0.898	0.967
Goddard <i>et al.</i> [6]	4.935	0.114	0.861	0.949	0.976
Gan <i>et al.</i> [9]	3.933	0.098	0.890	0.964	0.985
DORN [2]	2.727	0.072	0.932	0.984	0.994
Yin <i>et al.</i> [23]	3.258	0.072	0.938	0.990	<u>0.998</u>
BTS [3]	2.756	<u>0.059</u>	0.956	0.993	<u>0.998</u>
DPT-Hybrid [18]	2.573	0.062	0.959	0.995	0.999
Adabins [19]	<u>2.360</u>	0.058	<u>0.964</u>	<u>0.995</u>	0.999
Depthformer (ours)	2.285	0.058	0.967	0.996	0.999

Table 2. Results on KITTI Dataset. The best results are in **bold** and second best are underlined.

Method	Reference	Param. (M)	$RMSE \downarrow$	$REL \downarrow$
ResNet-50 [27]	<i>CVPR'16</i>	23.5	0.510	0.152
PVTv1 [12]	<i>ICCV'21</i>	23.9	0.508	0.166
Swin-T [15]	<i>ICCV'21</i>	27.5	0.456	0.142
Twins-SVT-S [16]	<i>NeurIPS'21</i>	23.5	0.443	0.141
MiT-B2 [11]	<i>NeurIPS'21</i>	24.2	0.394	0.118
MPVIT [17]	<i>CVPR'22</i>	22.6	0.403	0.120

Table 3. Performance of different state of the art multiscale-vision transformers for monocular depth estimation on the benchmark NYUV2 dataset.

across all the metrics for both outdoor scenes dataset KITTI and indoor dataset NYUV2. Our method reduces the $RMSE$ by 3.5% for NYUV2 against SoTA DPT-Hybrid [18] and by 3.2% against Adabins [19]. As shown in Fig. 2, our method Depthformer predicts a more accurate depth image in comparison to SoTAs [18] and [19].

Benchmarking of ViTs on Monocular depth estimation:

Table 3 shows the performance of various SoTA ViTs for monocular depth estimation on the NYUV2 dataset. For a fair comparison, we have taken model variants with a similar number of trainable parameters. For this experiment, we have taken the decoder as in [11]. All the ViTs outperform the CNN-based Resnet-50 [27]. MiT-B2 [11] shows the best performance in comparison to other ViTs. This can be attributed to shared information across feature vectors as mentioned in Section 2. In this work, we initialise the encoder via pretrained weights from MiT-B4 [11]. We hope this benchmark can be used as a baseline for future research in depth estimation.

Method	$RMSE \downarrow$	$REL \downarrow$
Decoder Xie <i>et al.</i> ([11])	0.375	0.114
Decoder(ours) + GAP	0.350	0.105
Decoder(ours) + Adabins [19]	0.394	0.115
Decoder(ours) + Transbins	0.345	0.100

Table 4. Ablation study on NYUV2 dataset for different decoder designs in the proposed Depthformer model.

Ablation for Decoder architecture: Table 4 showcases the efficacy of our proposed decoder network. For all these experiments, we have taken the proposed encoding backbone. The baseline uses a decoder network as proposed in [11], which predicts the depth at a resolution of $1/4th$ scale and upsamples it using bilinear interpolation. We first deploy Global Average Pooling (GAP) on top of our decoder to predict adaptive bins. As shown in the table, adding GAP gives a significant improvement over the baseline. Next, we ablate on our decoder design using the adaptive binning strategy proposed by Adabins [19] and our proposed Transbins. As shown, Transbins outperforms Adabins by a good margin which validates our claim that encoding global information during the initial layers is a better strategy in comparison to encoding the information at the final layers.

4. CONCLUSION

This paper introduces a multiscale vision transformer-based monocular depth estimation technique that achieves state-of-the-art results on KITTI and NYUV2 datasets. We also benchmark SoTA transformer architectures for monocular depth estimation task. We hope it motivates researchers to design new architectures to address task specific intricacies.

Acknowledgement: We acknowledge National Supercomputing Mission (NSM) for providing computing resources of 'PARAM Siddhi-AI', under National PARAM Supercomputing Facility, C-DAC Pune, and supported by the Ministry of Electronics and Information Technology and Department of Science and Technology, Government of India. This work has also been partly supported by the funding received from DST through the IMPRINT program (IMP/2019/000250).

5. REFERENCES

- [1] David Eigen, Christian Puhrsch, and Rob Fergus, “Depthmap Prediction From A Single Image Using A Multi-scale Deep network,” in *NIPS 2014*.
- [2] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao, “Deep Ordinal Regression Network For Monocular Depth Estimation,” in *CVPR 2018*.
- [3] 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,” .
- [4] Lam Huynh, Phong Nguyen-Ha, Jiri Matas, Esa Rahtu, and Janne Heikkilä, “Guiding Monocular Depth Estimation Using Depth-Attention Volume,” in *ECCV 2020*.
- [5] Xiaotian Chen, Xuejin Chen, and Zheng-Jun Zha, “Structure-Aware Residual Pyramid Network For Monocular Depth Estimation,” in *IJCAI-19*.
- [6] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic, “Unsupervised Video Summarization With Adversarial LSTM Networks,” in *CVPR 2017*.
- [7] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid, “Learning Depth From Single Monocular Images Using Deep Convolutional Neural Fields,” *ITPAMI 2016*.
- [8] Yevhen Kuznetsov, Jorg Stuckler, and Bastian Leibe, “Semi-Supervised Deep Learning For Monocular Depth Map Prediction,” in *CVPR 2017*.
- [9] Yukang Gan, Xiangyu Xu, Wenxiu Sun, and Liang Lin, “Monocular Depth Estimation With Affinity, Vertical Pooling, And Label Enhancement,” in *ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, Eds.
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in *CVPR 2009*.
- [11] Enze Xie, “SegFormer: Simple and Efficient Design For Semantic Segmentation With Transformers,” in *NIPS 2021*.
- [12] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao, “Pyramid Vision Transformer: A Versatile Backbone For Dense Prediction Without Convolutions,” in *ICCV 2021*.
- [13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, “An Image Is Worth 16x16 Words: Transformers For Image Recognition At Scale,” in *ICLR 2021*.
- [14] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, “End-to-End Object Detection with Transformers,” in *ECCV 2020*.
- [15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows,” in *ICCV 2021*.
- [16] Xiangxiang Chu, “Twins: Revisiting The Design Of Spatial Attention In Vision Transformers ,” in *NIPS 2021*.
- [17] Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang, “MPViT: Multi-Path Vision Transformer For Dense Prediction,” in *CVPR 2022*.
- [18] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun, “Vision Transformers For Dense Prediction,” in *ICCV 2021*.
- [19] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka, “AdaBins: Depth Estimation Using Adaptive Bins,” in *CVPR 2021*.
- [20] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, “Vision Meets Robotics: The KITTI Dataset,” *IJRR 2013*.
- [21] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, “Indoor Segmentation And Support Inference From RGBD Images,” in *ECCV 2012*.
- [22] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, “Feature pyramid networks for object detection,” in *CVPR 2017*.
- [23] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan, “Enforcing Geometric Constraints Of Virtual Normal For Depth Prediction,” in *ICCV 2019*.
- [24] Ravi Garg, Vijay Kumar B.G., Gustavo Carneiro, and Ian Reid, “Unsupervised CNN For Single View Depth Estimation: Geometry To The Rescue,” in *ECCV 2016*.
- [25] Ilya Loshchilov and Frank Hutter, “Decoupled Weight Decay Regularization,” in *ICML 2019*.
- [26] Leslie N. Smith and Nicholay Topin, “Super-Convergence: Very Fast Training of Residual Networks Using Large Learning Rates,” *CoRR*.
- [27] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell, “Deep Compositional Captioning: Describing Novel Object Categories Without Paired Training Data,” in *CVPR 2016*.