IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2023, London, UK. Cite as RA-L paper.

Detaching and Boosting: Dual Engine for Scale-Invariant Self-Supervised Monocular Depth Estimation

Peizhe Jiang, Wei Yang, Xiaoqing Ye, Xiao Tan, and Meng Wu

***Abstract*—Monocular depth estimation (MDE) in the self-supervised scenario has emerged as a promising method as it refrains from the requirement of ground truth depth. Despite continuous efforts, MDE is still sensitive to scale changes especially when all the training samples are from one single camera. Meanwhile, it deteriorates further since camera movement results in heavy coupling between the predicted depth and the scale change. In this paper, we present a scale-invariant approach for self-supervised MDE, in which scale-sensitive features (SSFs) are detached away while scale-invariant features (SIFs) are boosted further. To be specific, a simple but effective data augmentation by imitating camera zooming process is proposed to detach SSFs, making the model robust to scale changes. Besides, a dynamic cross-attention module is designed to boost SIFs by fusing multi-scale cross-attention features adaptively. Extensive experiments on the KITTI dataset demonstrate that the detaching and boosting strategies are mutually complementary in MDE and our approach achieves new State-of-The-Art performance against existing works from 0.097 to 0.090 w.r.t absolute relative error. The code will be made public soon.**

***Index Terms*—Autonomous Vehicle Navigation, Deep Learning for Visual Perception, Monocular Depth Estimation**

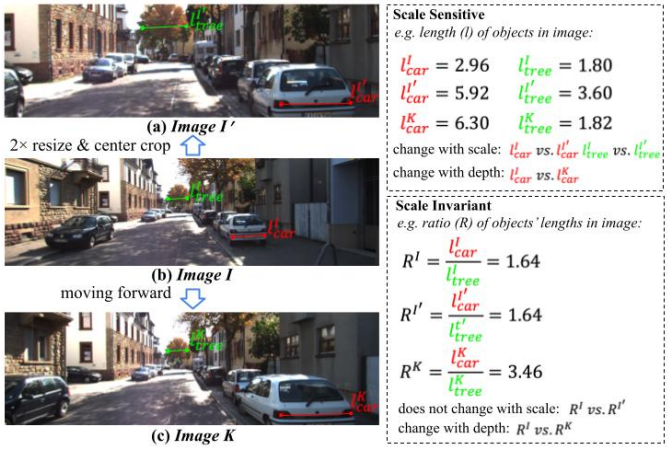

Fig. 1. Illustration on the influence of scale changes to the features. Image I' is an enlarged version of Image I imitating the scene captured by the camera with a larger focal length from the same viewpoint, and hence the depths of the same object in both Image I and Image I' should be the same. While Image K is captured after the camera moves forward from Image I, the depths of the same object in Image I and Image K are different.

I. INTRODUCTION

MONOCULAR depth estimation (MDE) is a critical but challenging computer vision (CV) task, which has a wide range of applications in augmented reality and autonomous driving. With the surge of convolutional neural networks (CNN), most supervised approaches [1], [2] have achieved leading performance. Nevertheless, ground truth (GT) of depth annotations is labor-intensive due to data sparsity and depth-sensing devices cost.

Recently, self-supervised approaches have gained more attention, because significant progress has been made on unsupervised learning of depth and ego-motion from unlabeled monocular video [3], [4], stereo pairs [5], [6] or a combination of both. The task tends to predict the depth by exploiting the geometrical relations between target and source views in the training data. Despite the less requirement for data preparation, self-supervised approaches are still striving for a decent performance, especially when compared with their supervised counterparts. Most of previous works made large efforts to address limitations of the photometric difference by adding further auxiliary constraints [4], [7], [8] since the supervision from view synthesis presents a dilemma, meaning strict pixel-wise correspondence between views.

Different from aforementioned works, we find that the change of depth is often accompanied by the scale change in camera movement, which is believed to be the very reason for severe performance degradation in MDE. By the term "scale", we mean the visual cues related with the appearance of an object in the image, which is determined jointly by the distance, camera focal length and the physical size of the object. To better illustrate this problem in Fig. 1, we coarsely classify the visual cues for depth estimation into scale-sensitive features (SSFs) and scale-invariant features (SIFs). SSFs refer to the features which are easily influenced by both scale and depth in images, such as the pixel-wise length of the object in images. To put it vividly, the tree and car in Fig. 1 (a) appear larger in size and closer in distance than those in Fig. 1 (b) while they come from the same picture actually, which indicates the estimated depths in both cases should be

Manuscript received: May 11, 2022; Revised July 19, 2022; Accepted September 17, 2022.

This paper was recommended for publication by Editor C. Cadena Lerma upon evaluation of the reviewers' comments. This work was supported by the Natural Science Basic Research Plan in Shaanxi Province of China (Grant No. 2020JQ-208). *(Corresponding author: Meng Wu.)*

Peizhe Jiang and Meng Wu are with the school of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: pz.jiang@mail.nwpu.edu.cn; wumeng@nwpu.edu.cn)

Wei Yang, Xiaoqing Ye and Xiao Tan are with Department of Computer Vision Technology (VIS), Baidu Inc., (e-mail: well_young@163.com; yexiaoqing@baidu.com; tanxchong@gmail.com)

Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2023, London, UK. Cite as RA-L paper.

equivalent in theory. On the contrary, other features which remain constant against the scale change but merely vary with depth are termed as scale-invariant features (SIFs), such as the ratio of pixel-wise lengths of two objects in the image. In Fig. 1, R^I and $R^{I'}$ are equal as they are from the same image; while R^K is different since the objects in image I and image K do have different depths. As is pointed out, this key observation motivates us to extract more SIFs but less SSFs for robust and reliable depth estimation. To fulfill the goal, we first resort to a novel method of camera zoom data augmentation (CZDA) to detach SSFs as much as possible, enabling the network to focus more on SIFs. After detaching SSFs, SIFs can be further enhanced via a dynamic cross-attention (DCA) module, which makes full use of multi-scale features by cross-attention and dynamic fusion.

To summarize, our contributions are three-folds as follows:

- The naive concepts of SSFs and SIFs are introduced for the first time in self-supervised monocular depth estimation.
- A SIF-based MDE method consisting of detaching and boosting is proposed in a plug-and-play manner for various self-supervised frameworks. The experiments indicate the effectiveness and complementarity of these two measures.
- We achieve new SOTA performance on the KITTI benchmark, which outperforms all the previous self-supervised MDE methods by a large margin.

II. RELATED WORK

A. Supervised Depth Estimation

Methods falling into this category require ground truth annotations of depth, which are usually cumbersome. Recent years have witnessed the prosperity of CNN since the breakthrough work of [9] achieves overwhelming performance over traditional methods. DORN [1] considers depth estimation as a classification task, in which depth discretization needs to be conducted at intervals for both GT and estimation. In another vein, Miangoleh et al. [10] proposed to boost existing MDE models by merging content-adaptive multi-scale estimations for high-resolution depth maps. To avoid over-fitting, several works [1], [9], [11] utilized scale-related data augmentation for depth estimation. Despite the continuous and thriving progress made so far, supervised methods are inherently defamed in the thirst for a huge amount of densely annotated data, which is a formidable task compared with other CV tasks.

B. Self-Supervised Depth Estimation

As an alternative, self-supervised methods learn to predict depth without labels, in which the self-supervision can be easily obtained by clues like geometric constraints between multiple views. Gary et al. [12] first applied a self-supervised training method with stereo pairs, in which photometric consistency loss using L_2 norm usually generates blurry depth maps. To solve this problem, Godard et al. [13] replaced it with SSIM and L_1 norm to yield better results. From then on, a variety of successive works have developed further in terms of objective functions [14], [15] or model architectures [5], [6].

Another stream of works predicts depth through camera ego-motion derived from monocular video, which is less demanding in data preparation but more challenging due to the unknown camera pose and moving objects. Many attempts have been made, including minimum reprojection loss [3] for dealing with occlusion, feature-metric loss [4] for enhancing textureless regions, and depth factorization and residual pose estimation for indoor environments [16]. In addition, other methods try to exploit semantic information from pre-trained module [17] or semantic labels [18] to enforce depth consistent over dynamic regions or near the object boundaries. When it comes to the problem of scale ambiguity, several works attempt to enforce scale consistency by formulating scale-consistent geometric constraints between multiple views [19], [20]. In this study, we aim to fully utilize scale-invariant features for depth estimation, based on a previously overlooked observation that severe coupling of the predicted depth with the scale change will result in degraded performance. Therefore, we seek a scale-invariant approach by detaching SSFs and boosting SIFs, which guides the network to predict depth robust to scale change.

C. Self-attention

The concept of attention started its dominance in natural language processing (NLP), and later in computer vision with its early success in CNN and later prosperity in Transformer. For depth estimation, Li et al. [21] combined with transformer to replace the original network for stereo matching. Johnston et al. [22] explored non-contiguous image regions as a context for estimating similar depth. Yan et al. [23] made the DepthNet to better understand the 3D structure of the whole scene and Jung et al. [17] designed a cross-task attention module for refining depth features more semantically consistent. Most self-attention based methods mentioned above excel in capturing long-range dependencies and aggregating discriminative features, which guarantees a steady performance increase. Bearing the advantage of self-attention in mind, we further enhance SIFs hierarchically by a newly-designed dynamic cross-attention.

III. METHOD

In this section, we first review the general framework for self-supervised monocular depth estimation, which consists of a DepthNet and a PoseNet. Then, we describe in detail how to integrate two core components of our SIF based approach, i.e., camera zoom data augmentation for detaching SSFs and dynamic cross-attention for boosting SIFs, into the whole framework seamlessly.

As is illustrated in Fig. 2, the image sequences are first transformed by camera zoom data augmentation before feeding to the DepthNet. It is assumed that SSFs can be eliminated to some degree by image zooming operation. Then a dynamic cross-attention module is involved to further encourage the network to predict depth from SIFs. Moreover, we leave PoseNet unchanged to get relative poses. With estimated depth

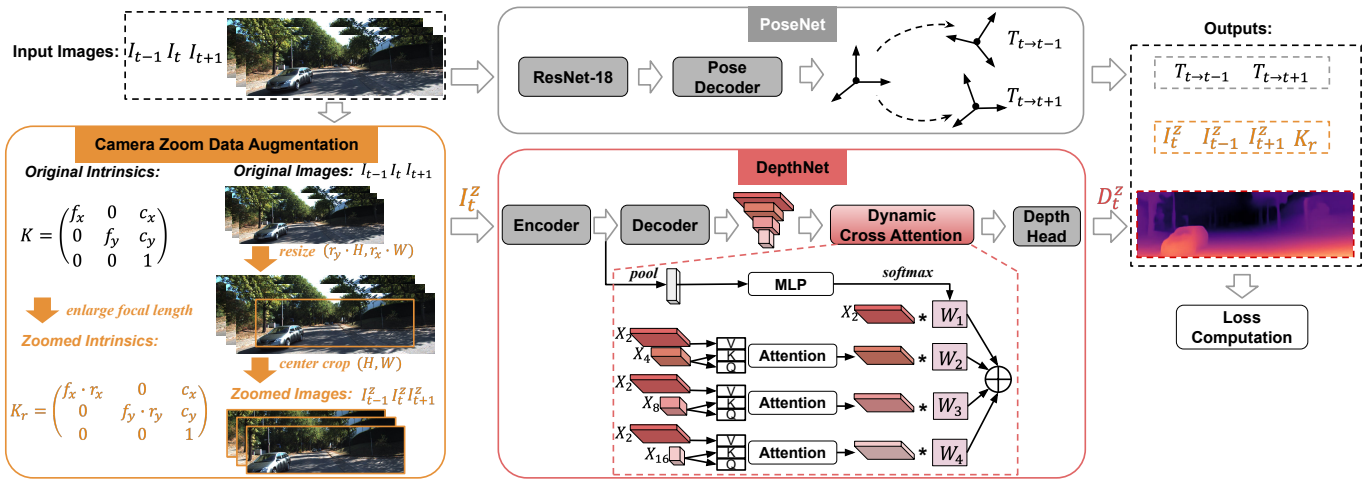

Fig. 2. The overview of our pipeline. First the input camera intrinsic matrix K and image sequence (I_{t-1}, I_t, I_{t+1}) are zoomed as K_r and $(I^z_{t-1}, I^z_t, I^z_{t+1})$ by camera zoom data augmentation. Then original image sequences are fed into the PoseNet to predict the relative poses $T_{t\to t-1}$ and $T_{t\to t+1}$. Taking the zoomed image I^z_t as input, DepthNet outputs the depth map D^z_t. Zoomed data, relative poses and predicted depth map are used for training PoseNet and DepthNet.

and relative camera pose, a reprojected target image from a source image can be reconstructed by pixel-wise correspondence and thus act as the self-supervision signal.

A. Self-supervised Monocular Depth Estimation

Given consecutive RGB images I_t and I_s, $s \in \{t-1, t+1\}$, and camera intrinsic matrix $K \in \mathbb{R}^{3\times3}$, we can derive the depth map D_t of I_t by DepthNet. Using monocular video as training, we need an auxiliary PoseNet to predict the relative pose $T_{t\to s}$ between I_t and I_s. Then we can warp the image I_s to I_t, following [24] by:

$$I_{s\to t} = I_s \left\langle K T_{t\to s} D_t K^{-1} p_t \right\rangle, \tag{1}$$

where p_t represents homogeneous pixel coordinates of I_t, and reconstruction process [24] is mainly used in $\langle\rangle$. $I_{s\to t}$ is the reprojected image from I_s to I_t. Following previous studies [3], [24], we stand on the foundation of structure from motion and construct the photometric loss L_{ph}, minimizing the discrepancy between I_t and $I_{s\to t}$ to optimize DepthNet and PoseNet. In detail, L_{ph} consists of the structural similarity index measure (SSIM) [25] and L_1 loss:

$$L_{ph}(I_t, I_{s\to t}) = \alpha \frac{1-SSIM(I_t, I_{s\to t})}{2} + \beta \left| I_t - I_{s\to t} \right|, \tag{2}$$

where SSIM computes over a 3×3 pixel window, with hyperparameters α and β. In addition, we apply minimum reprojection loss to deal with occlusions, and auto-mask to ignore static pixels where no relative camera motion is observed in monocular training [3].

Following [26], the edge-aware smoothness loss L_{sm} is also added to prevent shrinking of the estimated depth:

$$L^i_{sm} = \sum_p e^{-\left|\nabla^i I(p)\right|_1} \left| \nabla^i \hat{D}_t(p) \right|_1, \tag{3}$$

where $\hat{D}_t = D_t / \bar{D}_t$ is the mean-normalized inverse depth.

Finally, the loss function of our baseline is the combination of the reprojection loss and the smooth loss:

$$L = \lambda L_{ph} + \mu \left(L^1_{sm} + L^2_{sm} \right), \tag{4}$$

where λ and μ are used to balance their contributions.

B. Camera Zoom Data Augmentation for Detaching SSFs

As is demonstrated in Fig. 1, features are coarsely classified into SSFs and SIFs by our definition. For SSFs, the appearance features of the same instance under different depth ranges can vary conspicuously in monocular image sequences, leading to the strong correlation between depth estimation and object scales in the image. Inspired by the intuitive way humans perceive distant objects, we propose to mathematically adjust the focal length to decouple SSFs from the raw features and thus leave SIFs intact, which is the very essence of our camera zoom data augmentation. Specifically, the new data augmentation tries to imitate camera zooming where the focal length and resulting images are changed simultaneously such that the resulting depth is kept unchanged.

Suppose the original input of our pipeline is $I \in \mathbb{R}^{3\times H\times W}$, where H and W are the height and width of the original image. Camera intrinsic matrix is K with focal length f_x, f_y and principal point c_x, c_y. Here we modify K to get the zoomed intrinsic matrix K_r as follows:

$$K_r = K \cdot r = \begin{bmatrix} f_x \cdot r_x & 0 & c_x \\ 0 & f_y \cdot r_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{5}$$

where r_x and r_y denote the zoom ratio along width and height. Note that the ratios we adopt for each image are randomly generated with a given probability at each iteration. Correspondingly, we need to resize I to get image $I' \in \mathbb{R}^{3\times h\times w}$, where $h = H \cdot r_y$, $w = W \cdot r_x$, and then center-crop I' to the size of (H, W) with the principal point as the center so that the final image we use is the zoomed image $I^z \in \mathbb{R}^{3\times H\times W}$.

Since the focal length has been changed, the images need to be resized and center-cropped correspondingly to ensure that

the same 3D points in the world coordinate can correspond to the same pixels in I and I^z. To be specific, for a 3D point $S:[X,Y,Z]^T$ in camera coordinate system, $p:[u,v]^T$ is the pixel in image I corresponds to S, $p_r:[u_r,v_r]^T$ is the pixel in image I^z corresponds to S, so they have:

$$\begin{bmatrix} u_r - c_x \\ v_r - c_y \\ 1 \end{bmatrix} = \begin{bmatrix} f_x \cdot r_x & & \\ & f_y \cdot r_y & \\ & & 1 \end{bmatrix} \cdot \begin{bmatrix} X/Z \\ Y/Z \\ 1 \end{bmatrix} = \begin{bmatrix} (u - c_x) \cdot r_x \\ (v - c_y) \cdot r_y \\ 1 \end{bmatrix}. \quad (6)$$

Equation (6) is exactly the resize and center-crop process from p to p_r. It proves that our zoom process can keep the 3D geometric property unchanged, i.e., CZDA will not change the geometric relationship between objects and cameras.

The camera zoom process does not change the pose of adjacent frames so we use the original images as inputs of PoseNet, since the full image benefits to estimating pose by providing more context knowledge. Here (1) becomes:

$$I^z_{s\to t} = I^z_s \left\langle K_r T_{t\to s} D^z_t K_r^{-1} p^z_t \right\rangle, \quad (7)$$

where I^z_s is the zoomed I_s, D^z_t is the depth map of I^z_t, and p^z_t is pixel coordinates of I^z_t. So the loss function in (4) is computed between I^z_t and $I^z_{s\to t}$.

Based on the above deduction, it is assumed that the estimated depth is the unique factor influencing the reconstruction loss in (2) when feeding the data generated by CZDA. As a 3D geometry-driven method, the DepthNet is enforced to predict the same depth for different zoomed images so that the phtotometric loss in (2) can converge. In other words, SSFs can be detached from SIFs as much as possible during the training process, because the DepthNet is guided to predict the depth with more scale-insensitive features. Note that our CZDA is quite different from previous scale-related augmentation [9], [11] applied in the supervised methods, which changes the depth according to the zoom ratio. It is indicated that in self-supervised methods traditional data augmentation will break the strict pixel correspondence between views, thereby worsening the performance [27]. After detaching SSFs in this way, SIFs play a dominant role for DepthNet, and the predicted depth is thereby believed to be more robust to scale change.

C. Dynamic Cross-attention for Boosting SIFs

After detaching SSFs by CZDA, we expect to further boost SIFs, rendering the network more robust to scale change. As observed in previous works [23], [28], both attention models and feature fusion strategies across various scales are beneficial to construct more reliable features for depth estimation. Therefore, we propose a dynamic cross-attention module which takes full advantage of multi-scale features by cross-attention and dynamic fusion.

Given the feature maps X_2, X_4, X_8 and X_{16} from the decoder of DepthNet in Fig. 3 (a), X_i, $i \in [4,8,16]$ is supposed to be relatively global in a hierarchical way. Inspired by classic self-attention formula [29], we then define cross-attention as follows:

$$A(X_i, X_2) = X_2 + softmax\left(Q(X_i)^T K(X_i)\right) V(X_2), \quad (8)$$

where the query $Q(\cdot)$, key $K(\cdot)$, and value $V(\cdot)$ represent 1×1 convolution preceded by pooling and followed by

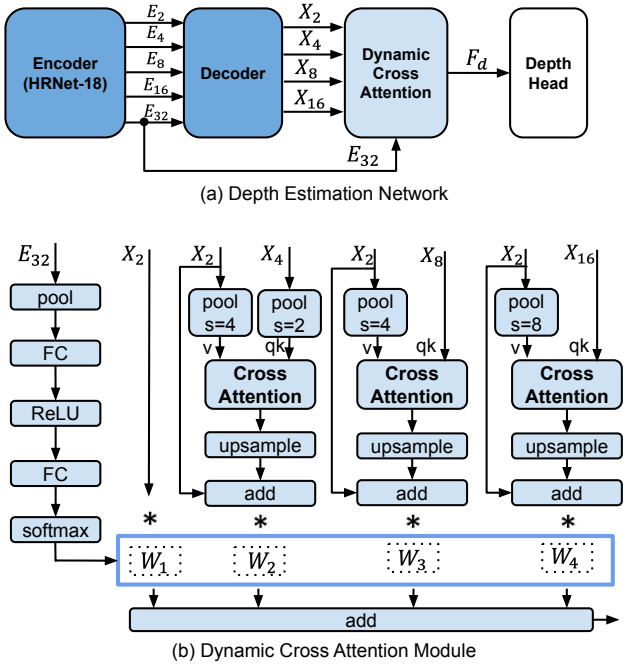

Fig. 3. Detail of our dynamic cross-attention. The encoder and decoder structure is the same as DIFFNet.

up-sampling to facilitate computation. The attention map $softmax\left(Q(X_i)^T K(X_i)\right)$ in Eq. (8) indicates the structural self-similarity of X_i. To obtain cross-attention features hierarchically, we feed the multi-scale features, like X_4, X_8 and X_{16}, to Q and K of cross-attention blocks correspondingly. For V, X_2 is kept constant. Pooling with adaptive stride beforehand is used to ensure the same feature dimension of two inputs for cross-attention block while up-sampling afterward is to match the dimension of X_2. As i varies from 4 to 16, X_i becomes more global and thus the cross-attention features $A(X_i, X_2)$ capture more coarse-grained structure, but less fine-grained details.

With these cross-attention features, our next concern is how to fuse them for enhanced SIFs. Here, we follow [30] to generate dynamic weights for feature fusion indicated by:

$$W = softmax(MLP(GAP(E_{32}))), \quad (9)$$

where $W \in \mathbb{R}^{1\times 4}$, E_{32} is the highest level feature of the encoder, MLP is multi-layer perception, and GAP stands for global average pooling. The final enhanced SIFs of F_d are then defined as:

$$\begin{aligned} F_d = {} & W_1 X_2 + W_2 A(X_4, X_2) + W_3 A(X_8, X_2) \\ & + W_4 A(X_{16}, X_2). \end{aligned} \quad (10)$$

With F_d, a depth head is added to predict the depth. We expect dynamic fusion to focus more on cross-attention features of coarse-grained structure for images in larger context since their layout is more important to depth estimation. Meanwhile, dynamic fusion is supposed to excite more fine-grained structural features in smaller context instead.

TABLE I

COMPARISON RESULTS ON THE KITTI DATASET. BEST RESULTS ARE IN **bold**, SECOND BEST ARE UNDERLINED. FOR red METRICS, LOWER IS BETTER, AND HIGER IS BETTER FOR blue METRICS. ABBREVIATION IN DATA COLUMN: D REFERS TO SUPERVISED METHODS WITH GT DEPTH SUPERVISION, $D^{\dagger}$ USES AUXILIARY DEPTH SUPERVISION FROM SLAM, D^{*} USES AUXILIARY DEPTH SUPERVISION FROM SYNTHETIC DEPTH LABELS, +SEM MEANS ADDITIONAL SUPERVISION FROM SEMANTIC LABELS OR PRE-TRAINED SEGMENTATION NETWORK, S REFERS TO TRAINING ON STEREO IMAGES AND M FOR TRAINING BY MONOCULAR VIDEOS. PP REPRESENTS POST-PROCESSING [13] FOR MOST METHODS.

Method	PP	Data	$H \times W$	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen [9]		D	184×612	0.203	1.548	6.307	0.282	0.702	0.890	0.890
Yang [31]	✓	$D^{\dagger}S$	256×512	0.097	0.734	4.442	0.187	0.888	0.958	0.980
Luo [32]		$D^{*}DS$	192×640 crop	0.094	0.626	4.252	0.177	0.891	0.965	0.984
DORN [1] RestNet		D	385×513 crop	0.072	0.307	2.727	0.120	0.932	0.984	0.994
AdaBins [2]		D	352×704	**0.058**	**0.190**	**2.360**	**0.088**	**0.964**	**0.995**	**0.999**
Monodepth [13]	✓	S	256×512	0.138	1.186	5.650	0.234	0.813	0.930	0.969
Monodepth2 [3]	✓	S	320×1024	0.105	0.822	4.692	0.199	0.876	0.954	0.977
DepthHints [14]	✓	S	320×1024	0.096	0.710	4.393	0.185	0.890	0.962	0.981
SingleNet [6]	✓	S	320×1024	0.094	0.681	4.392	0.185	0.892	0.962	0.981
FALnet [8]	✓	S	192×640 crop	0.094	0.597	4.005	0.173	0.900	**0.967**	**0.985**
EPCNet [27]	✓	S	320×1024	0.091	0.646	4.207	0.176	**0.901**	0.966	0.983
EdgeDepth [15]	✓	S+Sem	320×1024	0.091	0.646	4.244	0.177	0.898	0.966	0.983
PLADE-Net [5]		S	192×640 crop	0.092	0.626	4.046	0.175	0.896	0.965	0.984
PLADE-Net [5]	✓	S	192×640 crop	**0.089**	**0.590**	**4.008**	**0.172**	0.900	**0.967**	**0.985**
Monodepth2 [3]	✓	M	320×1024	0.112	0.838	4.607	0.187	0.883	0.962	0.982
Johnston. [22]		M	192×640	0.106	0.861	4.699	0.185	0.889	0.962	0.982
HR-Depth [33]		M	384×1280	0.104	0.727	4.410	0.179	0.894	0.966	0.984
Featdepth [4]		M	320×1024	0.104	0.729	4.481	0.179	0.893	0.965	0.984
CADepth-Net [23]		M	320×1024	0.102	0.734	4.407	0.178	0.898	0.966	0.984
FSRE-Depth [17]	✓	M+Sem	192×640	0.102	0.675	4.393	0.178	0.893	0.966	0.984
PackNet-SfM [34]	✓	M+Sem	320×1024	0.100	0.761	4.270	0.175	0.902	0.965	0.982
DIFFNet [35]		M	320×1024	0.097	0.722	4.345	0.174	0.907	0.967	0.984
Ours		M	320×1024	0.092	0.640	4.208	0.169	0.909	0.969	**0.985**
Ours	✓	M	320×1024	**0.090**	**0.597**	**4.087**	**0.167**	**0.912**	**0.970**	**0.985**

In fact, the statistical distribution of dynamic weights reveals that the weights W_1 and W_2 tend to be larger for a zoomed image, while W_3 and W_4 turn bigger for an original image, verifying the mechanism of DCA exactly. Due to the page limit, we will not provide the results in the paper. To elaborate it further, cross-attention generates multi-scale representation of the input features with hierarchical structure clue, and then the weights of the multi-scale features are dynamically adjusted for scale invariant features (SIFs) representation.

As two measures of CZDA and DCS are orthogonal in nature, they are supposed to be complementary. The experiments prove that enhancing SIFs in this way is necessary for MDE after SSFs are detached.

IV. EXPERIMENTS

In this section, we first make a comprehensive comparison with tens of SOTA methods. Then, we analyze the role of detaching SSFs, i.e., the significance of SIF's introduction. Finally, we perform extensive ablation studies on each component of our approach and validate the effectiveness and generalization of camera zoom data augmentation in other frameworks.

A. Dataset and Performance Metrics

Our model was trained on the KITTI 2015 dataset [36], which contains videos in 200 street scenes captured by RGB cameras, with sparse depth ground truths captured by Velodyne laser scanner. For training, we remove static frames by a pre-processing step suggested by [24]. We adopt the Eigen split of [9] to divide KITTI raw data, resulting in 39,810 monocular triplets for training, 4,424 for validation and 697 for testing.

The performance is assessed by standard metrics, like absolute relative difference, square related difference, RMSE and log RMSE. For the methods trained on monocular videos, the depth is defined up to scale factor $\hat{s}$ during evaluation [24], which is computed by:

$$\hat{s} = median\left(D_{gt}\right) / median\left(D_{pred}\right). \tag{11}$$

where the predicted depth maps D_{pred} are multiplied by the computed scale factor $\hat{s}$ to match the median of the ground truth D_{gt}. The scale factor $\hat{s}$ here is also an important inspector of SSFs, and we will discuss it later.

B. Implementation Details

Our DepthNet has a similar network architecture as DIFFNet [35], which adopts HR-net [37] as the encoder and the channel-wise attention in the decoder. Unlike DIFFNet that uses multi-scale depth maps at training, we predict a single depth map with the same scale of input for loss computation in (4). In terms of PoseNet, we simply follow the architecture in [4], considering the relative camera poses are fairly simple in outdoor environments (e.g., KITTI) [16]. DepthNet and PoseNet respectively take the image sizes of 320×1024 and 192×640 as inputs.

We follow the setting in [3] for data pre-processing. The models are trained and tested on 4 Tesla V100. Adam optimizer with the default betas 0.9 and 0.999 is used. The learning rate starts from $1e^{-4}$, and then decays to half of the original at the 20th and 30th epoch [4].

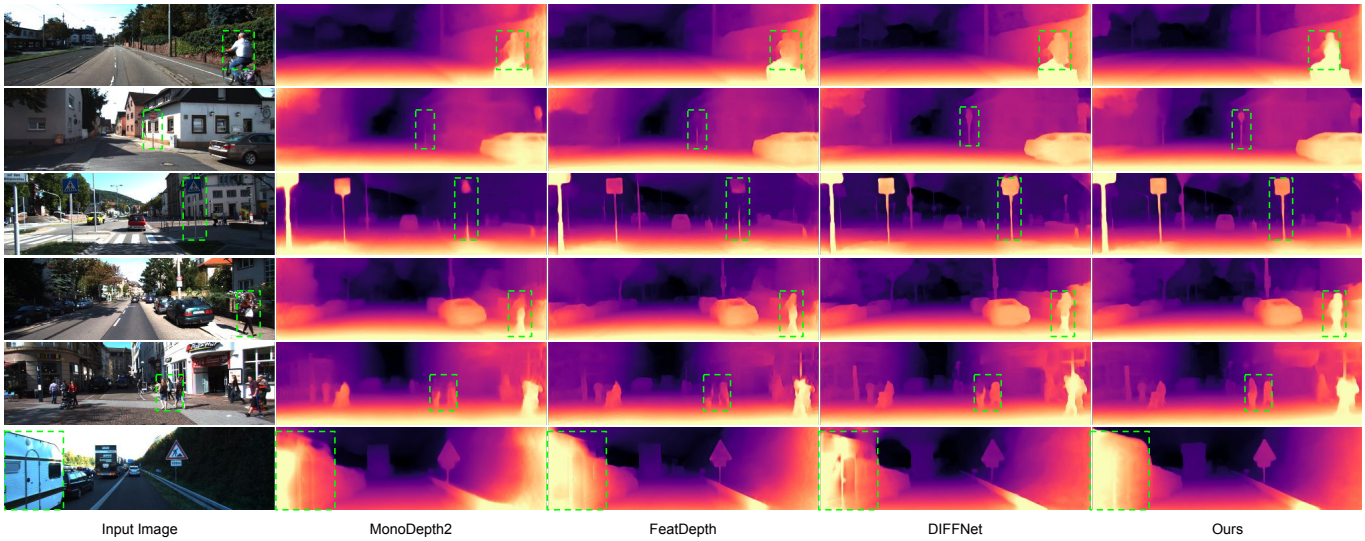

Fig. 4. Qualitative visualization results on KITTI Eigen Test Split without post-processing. Our approach performs better on thinner objects such as signs and bollards, as well as being better at delineating difficult object boundaries like poles and humans.

Unlike prior post-processing (PP) [5] [13], we here adopt a novel zoom average post-processing, which fuses different depth estimation results from different zoom scales. We run the depth estimator three times to get three different depth maps, one for the original image and the other two for $1.33\times$ and $2\times$ zoomed images respectively. To be specific, the images are first resized to the sizes of $(1.33H, 1.33W)$ and $(2H, 2W)$, and then center-cropped to (H, W) to get $1.33\times$ and $2\times$ zoomed images. Finally, we restore these depth maps to the original sizes by bilinear interpolation and average them directly in the corresponding region for the final output.

C. Evaluation on KITTI

Tab. I lists the quantitative results of various methods on the KITTI dataset [38]. During our evaluation, we cap depth to 80m per standard practice [13]. Note that FAL-Net and PLADE-Net adopt multi-scale post-processing (PP) strategy [5], while our approach utilizes the proposed zoom average PP measure for better results. The table is split into several parts according to the supervision level, the post-processing, and the size of images. The results in the bottom block of Tab. I show that two variants of our approach significantly outperform all existing SOTA self-supervised methods using monocular video w.r.t. all the metrics. Higher results can be achieved with our zoom average PP, which verifies the superiority of our approach in aggregating the depths of images with different scales. We also outperform recent methods with semantic label assisted [17] [34]. Furthermore, our approach yields comparable or better results than all of self-supervised methods using stereo images. As is known, post-processing strategies generally lead to the growth of running time by fusing different depth results. Our zoom average PP will increase about $3\times$ the runtime, as it fuses three different depth estimation results.

In addition, Fig. 4 illustrates the qualitative performance of our method against Monodepth2 [3], Featdeapth [4] and DIFFNet [35] without post-processing. It is obvious that our approach achieves better performance on thinner objects and low-texture regions, like signs, bollards, and the side of bus, while retaining finer details like silhouettes of humans and poles. Our approach presents overall more structurally consistent and accurate depth over the entire scene.

TABLE II
MEDIAN SCALE FACTOR CHANGE WITH AND WITHOUT CAMERA ZOOM DATA AUGMENTATION (CZDA) ON KITTI TEST SET.

Scale Factor	Ours	Ours+CZDA	Monodepth2	Monodepth2+CZDA
$median(\hat{s}_c)$	28.886	29.863	34.152	31.447
$median(\hat{s}_z)$	52.536	30.424	63.672	31.126

D. Analysis on Detaching Scale-Sensitive Features

To validate the proposed detaching strategy, we resort to scale factor mentioned above. Self-supervised MDE merely gets relative depth, while scale factor $\hat{s}$ is used to adjust the predicted depth map to GT, indicating whether the method is insensitive to scale change. In other words, for both original image and zoomed image, scale factors of methods with camera zoom data augmentation are supposed to be constant.

We test Monodepth2 [3] and our method with/without CZDA for a qualitative comparison, as shown in Fig. 5. $\hat{s}_c$ corresponds to the scale factor of the center crop region, and $\hat{s}_z$ is that of zoomed image.

It is obvious that $\hat{s}_c$ has little difference with $\hat{s}_z$ for these two methods trained with CZDA, while the difference between $\hat{s}_c$ and $\hat{s}_z$ varies largely without CZDA. A similar conclusion can be drawn from the colorized depth maps, i.e., the center crop region color does not change when trained with camera zoom data augmentation. Furthermore, we perform this experiment on all the test set of KITTI with the results shown in Tab. II. Experimental results verify that the proposed camera zoom data augmentation enables the network to detach SSFs, making

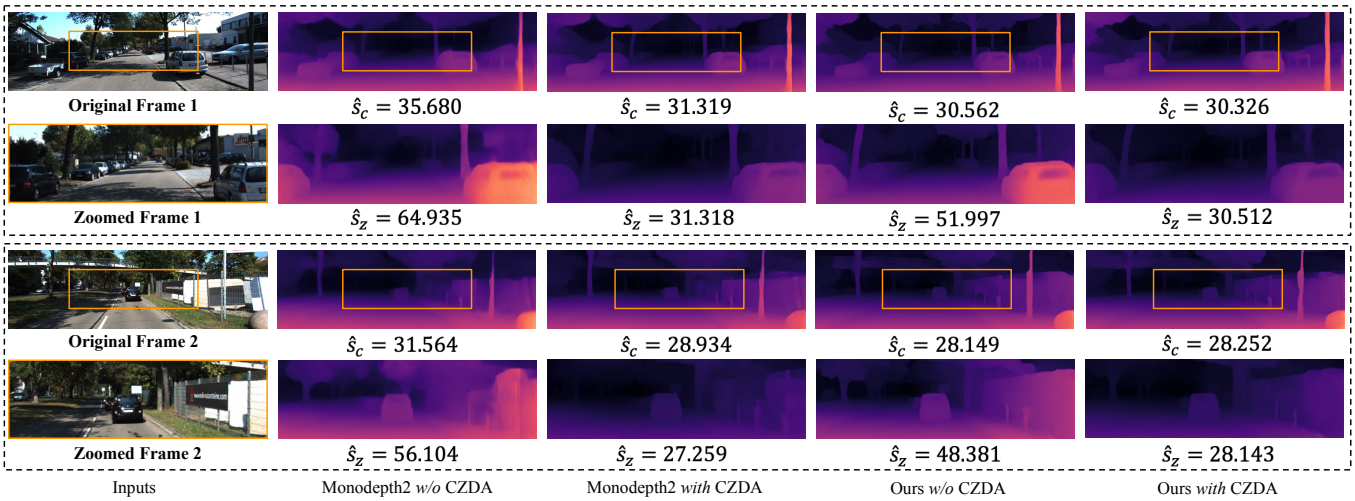

Fig. 5. Visualization of scale factor change. CZDA is the abbreviation of camera zoom data augmentation. For generating appropriate visualization results, minimum depth and maximum depth are fixed as 0.1 and 100 respectively for all the depth maps. Zoomed frames correspond to the orange center crop box region in original frames. Scale factor are computed as in (11), $\hat{s}_c$ is the scale factor corresponding to the orange center crop box, and $\hat{s}_z$ is the scale factor of zoomed image. The GT used for scale factor computation are the GT of center crop regions.

TABLE III
ABLATION RESULTS AMONGST VARIANTS OF OUR MODEL ON THE KITTI DATASET. CZDA REFERS TO CAMERA ZOOM DATA AUGMENTATION AND DCA REFERS TO DYNAMIC CROSS-ATTENTION MODULE.

CZDA	DCA	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
		0.096	0.676	4.280	0.173	0.905	0.968	0.984
✓		0.095	0.657	4.198	0.170	0.907	0.969	0.985
	✓	0.097	0.670	4.268	0.172	0.904	0.968	0.984
✓	✓	0.092	0.640	4.208	0.169	0.909	0.969	0.985

TABLE IV
RESULTS OF OUR MODEL WITH DIFFERENT ZOOM RANGES.

Zoom Range	Abs Rel	Sq Rel	RMSE	$\delta < 1.25$
1∼1.33	0.095	0.664	4.274	0.907
1∼2	0.092	0.640	4.208	0.909
1∼3.33	0.094	0.650	4.219	0.908
1.25∼3.33	0.094	0.663	4.228	0.907

TABLE V
RESULTS OF OUR CAMERA ZOOM DATA AUGMENTATION ON OTHER SELF-SUPERVISED MDE METHODS ON THE KITTI DATASET. CZDA REFERS TO CAMERA ZOOM DATA AUGMENTATION.

Method	Abs Rel	Sq Rel	RMSE	$\delta < 1.25$
Monodepth2	0.115	0.882	4.701	0.879
Monodepth2+CZDA	0.106	0.789	4.537	0.891
Featdepth	0.104	0.729	4.481	0.893
Featdepth+CZDA	0.098	0.662	4.331	0.901

TABLE VI
COMPARISON RESULTS OF VARIOUS ATTENTION MODULES ON THE KITTI DATASET. CA REFERS TO CHANNEL-WISE ATTENTION, SA REFERS TO SPATIAL-ATTENTION, AND CROSS-A REFERS CROSS-ATTENTION.

Method	Abs Rel	Sq Rel	RMSE	$\delta < 1.25$
CA [23]	0.096	0.773	4.461	0.906
SA [35]	0.095	0.687	4.265	0.907
Cross-A	0.095	0.669	4.236	0.907
DCA (ours)	0.092	0.640	4.208	0.909

the model robust to scale changes, and thus with our designed PP strategy, we can fuse different scales depths directly.

E. Ablation Study

In Tab. III, we present a comprehensive ablation study of the key components of the proposed approach on KITTI. The addition of camera zoom data augmentation brings consistent performance increase w.r.t all metrics. Applying dynamic cross-attention module further improves the performance. It is interesting that directly applying DCA to the baseline yields no positive effect. The reason behind is that DCA plays the role of deeply excavating the initial features. Without CZDA for detaching SSFs, the network with DCA is still prone to predict depths from SSFs. On the other hand, the combination of CZDA and DCA gains a large performance improvement, which proves that the two components are complementary.

To investigate CZDA further, we examine the key factor of focal length. Here we set our camera zoom data augmentation with different zoom ranges i.e., the random range of r_x and r_y in (5) and then validate our model with different settings. The results in Tab. IV show that the zoom range from 1 to 2 works best. One possible reason is that a wider range will bring more quantization error while a narrower range undermines the potential of CZDA.

We also apply our camera zoom data augmentation in other frameworks [3], [4], as is listed in Tab. V. The results demon-

strate the effectiveness and generalization of the proposed CZDA, which can be utilized in a plug-and-play way for self-supervised MDE and bring considerable improvement. Considering the training efficiency, the loss function is computed on only one scale for Featdepth [4].

Besides, we replace our DCA with other attention modules and give the comparison results on KITTT in Tab. VI. It's obvious that our DCA achieves the best performance, which is attributed to the dynamic fusion of cross-attention features.

V. CONCLUSION

In this paper, we bring forward a novel scale-invariant feature (SIF) based monocular depth estimation with a view to the observation that scale change matters. The proposed method consists of two measures of detaching and boosting. When it refers to detaching, the method invents a new data augmentation of camera zooming to detach SSF. Meanwhile, SIF is further boosted via a dynamic cross-attention module. Thorough experiments validate SIF and also indicate its striking superiority. In the future, we will explore the consistency constraint between depth maps from original image and zoomed image to improve the performance.

REFERENCES

[1] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *CVPR*, 2018, pp. 2002–2011.

[2] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *CVPR*, 2021, pp. 4009–4018.

[3] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *ICCV*, 2019, pp. 3828–3838.

[4] C. Shu, K. Yu, Z. Duan, and K. Yang, "Feature-metric loss for self-supervised learning of depth and egomotion," in *ECCV*, 2020, pp. 572–588.

[5] J. L. Gonzalez and M. Kim, "Plade-net: Towards pixel-level accuracy for self-supervised single-view depth estimation with neural positional encoding and distilled matting loss," in *CVPR*, 2021, pp. 6851–6860.

[6] Z. Chen, X. Ye, W. Yang, Z. Xu, X. Tan, Z. Zou, E. Ding, X. Zhang, and L. Huang, "Revealing the reciprocal relations between self-supervised stereo and monocular depth estimation," in *ICCV*, 2021, pp. 15 529–15 538.

[7] S. Lee, S. Im, S. Lin, and I. S. Kweon, "Learning monocular depth in dynamic scenes via instance-aware projection consistency," in *AAAI*, 2021, pp. 1863–1872.

[8] J. L. GonzalezBello and M. Kim, "Forget about the lidar: Self-supervised depth estimators with med probability volumes," *NeurIPS*, pp. 12 626–12 637, 2020.

[9] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *NeurIPS*, pp. 2366–2374, 2014.

[10] S. M. H. Miangoleh, S. Dille, L. Mai, S. Paris, and Y. Aksoy, "Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging," in *CVPR*, 2021, pp. 9685–9694.

[11] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *3DV*, 2016, pp. 239–248.

[12] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *ECCV*, 2016, pp. 740–756.

[13] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*, 2017, pp. 270–279.

[14] J. Watson, M. Firman, G. J. Brostow, and D. Turmukhambetov, "Self-supervised monocular depth hints," in *ICCV*, 2019, pp. 2162–2171.

[15] S. Zhu, G. Brazil, and X. Liu, "The edge of depth: Explicit constraints between segmentation and depth," in *CVPR*, 2020, pp. 13 116–13 125.

[16] P. Ji, R. Li, B. Bhanu, and Y. Xu, "Monoindoor: Towards good practice of self-supervised monocular depth estimation for indoor environments," in *ICCV*, 2021, pp. 12 787–12 796.

[17] H. Jung, E. Park, and S. Yoo, "Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation," in *ICCV*, 2021, pp. 12 642–12 652.

[18] M. Klingner, J.-A. Termöhlen, J. Mikolajczyk, and T. Fingscheidt, "Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance," in *ECCV*, 2020, pp. 582–600.

[19] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints," in *CVPR*, 2018, pp. 5667–5675.

[20] L. Wang, Y. Wang, L. Wang, Y. Zhan, Y. Wang, and H. Lu, "Can scale-consistent monocular depth be learned in a self-supervised scale-invariant manner?" in *ICCV*, 2021, pp. 12 727–12 736.

[21] Z. Li, X. Liu, N. Drenkow, A. Ding, F. X. Creighton, R. H. Taylor, and M. Unberath, "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *ICCV*, 2021, pp. 6197–6206.

[22] A. Johnston and G. Carneiro, "Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume," in *CVPR*, 2020, pp. 4756–4765.

[23] J. Yan, H. Zhao, P. Bu, and Y. Jin, "Channel-wise attention-based network for self-supervised monocular depth estimation," in *3DV*, 2021, pp. 464–473.

[24] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*, 2017, pp. 1851–1858.

[25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *TIP*, pp. 600–612, 2004.

[26] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey, "Learning depth from monocular videos using direct methods," in *CVPR*, 2018, pp. 2022–2030.

[27] R. Peng, R. Wang, Y. Lai, L. Tang, and Y. Cai, "Excavating the potential capacity of self-supervised monocular depth estimation," in *ICCV*, 2021, pp. 15 560–15 569.

[28] A. Sagar, "Monocular depth estimation using multi scale neural network and feature fusion," in *WACV*, 2022, pp. 656–662.

[29] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.

[30] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *CVPR*, 2020, pp. 11 030–11 039.

[31] N. Yang, R. Wang, J. Stückler, and D. Cremers, "Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry," in *ECCV*, 2018, pp. 817–833.

[32] Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, and L. Lin, "Single view stereo matching," in *CVPR*, 2018, pp. 155–163.

[33] X. Lyu, L. Liu, M. Wang, X. Kong, L. Liu, Y. Liu, X. Chen, and Y. Yuan, "Hr-depth: High resolution self-supervised monocular depth estimation," in *AAAI*, 2021, pp. 2294–2301.

[34] V. Guizilini, R. Hou, J. Li, R. Ambrus, and A. Gaidon, "Semantically-guided representation learning for self-supervised monocular depth," *arXiv preprint*, 2020.

[35] H. Zhou, D. Greenwood, and S. Taylor, "Self-supervised monocular depth estimation with internal feature fusion," in *BMVC*, 2021.

[36] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*, 2012, pp. 3354–3361.

[37] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *TPAMI*, pp. 3349–3364, 2020.

[38] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *ICCV*, 2015, pp. 2650–2658.