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

TransDSSL: Transformer based Depth Estimation via Self-Supervised Learning

Daechan Han[1*], Jeongmin Shin[1*], Namil Kim[2], Soonmin Hwang[3], and Yukyung Choi[1†]

***Abstract*—Recently, transformers have been widely adopted for various computer vision tasks and show promising results due to their ability to encode long-range spatial dependencies in an image effectively. However, very few studies on adopting transformers in self-supervised depth estimation have been conducted. When replacing the CNN architecture with the transformer in self-supervised learning of depth, we encounter several problems such as problematic multi-scale photometric loss function when used with transformers and, insufficient ability to capture local details. In this paper, we propose an attention-based decoder module, Pixel-Wise Skip Attention (PWSA), to enhance fine details in feature maps while keeping global context from transformers. In addition, we propose utilizing self-distillation loss with single-scale photometric loss to alleviate the instability of transformer training by using correct training signals. We demonstrate that the proposed model performs accurate predictions on large objects and thin structures that require global context and local details. Our model achieves state-of-the-art performance among the self-supervised monocular depth estimation methods on KITTI and DDAD benchmarks.**

***Index Terms*—Deep Learning for Visual Perception, Computer Vision for Transportation, Visual Learning**

I. INTRODUCTION

In recent years, monocular depth estimation has been actively researched in the computer vision and robotics fields due to its potential benefits, such as substituting expensive LiDAR sensors widely used in advanced robotic systems, including self-driving vehicles or enhancing other computer vision tasks performed in 3d space [28], [35]. However, a large-scale dataset with high diversity is generally required to train neural networks, and collecting enough data with accurate ground truth depth information for supervision is expensive and laborious, especially in outdoor environments. A promising way to alleviate this burden is to use self-supervised learning techniques based on synchronous stereo images [5] or monocular video [3], [4], [6] so that any supervision collected from an additional sensor, such as LiDAR is not required.

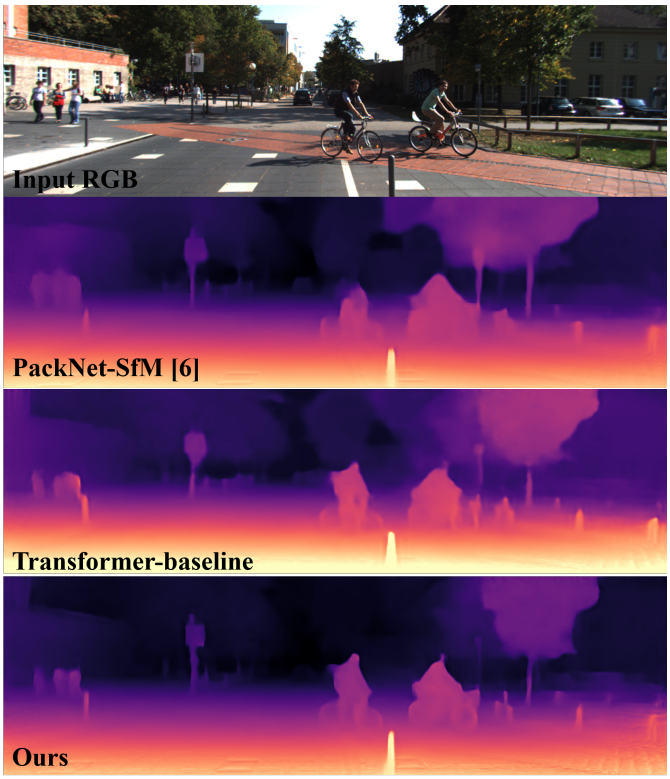


Fig. 1. **Example of depth predictions from a single image.** Compared to PackNet-SfM [6] prediction, "transformer-baseline", a naive approach by replacing the CNN encoders with a Swin-Transformer, results in a slight improvement. In addition, our final model shows further improvement such as sharp depth discontinuity around object boundaries because of the enhanced local details.

Even though the quality of the estimated depth from a monocular image has dramatically improved with the advancement of convolutional neural networks (CNNs) [2], [3], CNNs have been limited in modeling long-range spatial dependencies on many computer vision tasks [25], [27]. Many techniques have been proposed to overcome this limitation by encoding larger context, such as explicit modeling of the large receptive field by dilated [7] or deformable [19] convolution, employing self-attention mechanism [1], or aggregating multi-scale features [8]. Still, there is a fundamental limitation due to the locality of the convolution operation.

As a solution, a transformer [31], originally designed for natural language processing tasks to capture long-range dependencies in sequences, is adopted to various computer vision

Manuscript received: April 8, 2022; Revised: June 9, 2022; Accepted: July 19, 2022.

*This paper was recommended for publication by Editor Cesar Cadena Lerma and Editor Tamim Asfour upon evaluation of the Associate Editor and Reviewers' comments. This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020M3F6A1109603, 40%) and the MSIT (Ministry of Science and ICT), Korea, under the ICT Challenge and Advanced Network of HRD (RS-2022-00156345, 40%) supervised by the IITP (Institute for Information communications Technology Planning Evaluation), and Institute for Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea Government (MSIT) (No. 2021-0-02067, Next Generation AI for Multi-purpose Video Search, 20%) (*Daechan Han and Jeongmin Shin contributed equally to this work.)*

†*Corresponding author: Yukyung Choi*

[1]Daechan Han, Jeongmin Shin and Yukyung Choi are with School of Intelligent Mechatronic Engineering, Sejong University, South Korea {dchan,jmshin,ykchoi}@rcv.sejong.ac.kr

[2]Namil Kim is with NAVER LABS, South Korea namil.kim@naverlabs.com

[3]Soonmin Hwang is with Carnegie Mellon University, PA, USA soonminh@andrew.cmu.edu

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.

tasks, and it has shown promising results [29], [30], [32], [33]. However, the transformer could suffer from a lack of local details on a pixel-level task in which the fine details are important to make a good prediction [22], [38]. We conducted a preliminary "transformer-baseline" experiment, adopting the transformer instead of convolutional encoders on the self-supervised monocular depth estimation task. As shown in Fig. 1, it is limited in predicting fine details, such as thin structures like poles and the boundary of the objects, because of the lack of local information. We also observe another substandard optimizability issue [24], *i.e*, the "transformer-baseline" model has difficulty in training caused by noisy training signals from typical multi-scale photometric loss.

This paper presents a comprehensive study on training a better neural network with a transformer on a self-supervised monocular depth estimation task. We propose a novel Pixel-Wise Skip Attention (PWSA) module that encourages capturing local details while effectively keeping the global context from the transformer encoder. We also propose a novel self-distillation loss to replace the conventional multi-scale photometric loss. In training the depth network, the highest resolution of prediction is still supervised by single-scale photometric loss. However, our self-distillation loss leverages the highest resolution of prediction as a pseudo label to supervise other intermediate scales of prediction, which could enhance intermediate representation and the training stability. To minimize the error propagation from the incorrect pseudo label, we designed an adaptive weighting scheme for the multi-scale self-distillation loss. An interesting benefit from our self-distillation loss is that we can detach the last decoder part at inference time if the capability of predicting fine details is successfully transferred to the penultimate decoder after training. This helps improve the efficiency of the final model while maintaining accuracy.

II. RELATED WORK

A. Self-supervised Monocular Depth Estimation

Zhou *et al.* [12] proposed the framework for self-supervised training in the purely monocular setting, where a depth and pose network are simultaneously learned from unlabeled monocular videos. Godard *et al.* [3] proposed a minimum reprojection loss and auto-masking loss, to deal with occlusions and regions that violate static-world assumption by moving objects. By extending Godard *et al.* [3], there have been various methods that incorporate additional objectives and constraints as follows: Shu *et al.* [13] proposed the feature metric loss for textureless regions where the photometric loss has difficulty in the producing correct training signal. [6], [14] addressed the scale ambiguity problem of the monocular SfM-based method by simply using the instantaneous GPS or velocity of the camera during training. Guizilini *et al.* [6] also proposed a novel architecture with self-supervision to learn detail-preserving representations. Additionally, several works have utilized additional sources such as optical flow [15], segmentation label [16], and augmentation loss [17] to improve the quality of the depth prediction. While additional sources can improve accuracy, these sources require additional models and processing steps. In contrast, our self-distillation loss is a simple yet effective way of leveraging a detachable module to generate the pseudo label with fine details during training.

B. Global Context modeling

With the self-attention module, transformers can successfully catch global context information that CNNs have shown a fundamental limitation, so transformers have shown wide success on various computer vision tasks [9], [29], [32], [33]. ViT [20] was the first work to show that a pure transformer-based image classification model achieves state-of-the-art accuracy. Liu *et al.* [10] presented the Swin Transformer, a hierarchical transformer with a shifted windowing scheme to achieve an efficient network for object recognition tasks. In supervised monocular depth estimation, DPT [9] introduced a dense vision transformer that leverages ViT in place of the backbone model. TransDepth [18] proposed a hybrid model to incorporate the advantage of using a transformer and CNN simultaneously.

In contrast with the aforementioned works that are based on a supervised setting by explicit ground truth information, e.g., LiDAR supervision, Varma *et al.* [37] investigates a way to adopt vision transformers for self-supervised monocular depth estimation. However, it is shown that performance gain falls short of expectations contrary to the success stories on other vision tasks with a transformer. To successfully adopt the transformer for self-supervised learning, we propose a novel combination of a transformer and Pixel-Wise Skip Attention (PWSA) module that can generate fine-detailed and spatially coherent depth prediction.

III. PROPOSED METHOD

A. Review of Self-supervised Learning

The conventional training process of self-supervised monocular depth estimation from videos, established in Zhou *et al.* [12], is designed as a multi-task learning pipeline for estimating depth and motion between images simultaneously. Some common assumptions are that 1) the camera intrinsic matrix K is given, 2) the camera is moving over time, and 3) the world is static. During training, the depth network predicts a pixel-wise depth map D_t from a given input target image I_t. Pose network takes the target image I_t and a source image I_s as input and predicts a 6-DoF relative motion $T_{t\to s}$ between them. Then, the source image is warped to synthesize the target image $\hat{I}_t$ as shown in Eq. 1. Then photometric consistency loss $\mathcal{L}$ between the target image I_t and the synthesized target image $\hat{I}_t$ is calculated to update both depth and pose networks. In reality, auto-masking [9] is widely used to deal with moving objects in the scene which break the static world assumption. In addition, velocity loss [6] that constrains translation components in $T_{t\to s}$ can be applied.

$$\begin{aligned} &\hat{I}_t(\mathbf{x}) \mapsto I_s(G_{t\to s}(\mathbf{x})) \\ &\quad \text{where } G_{t\to s}(\mathbf{x}) = KT_{t\to s}(K^{-1}\mathbf{x} \odot D_t), \\ &\quad \mathbf{x} \text{ indicates pixel coordinates in image space,} \\ &\quad \text{and } \odot \text{ means element-wise multiplication.} \end{aligned} \tag{1}$$

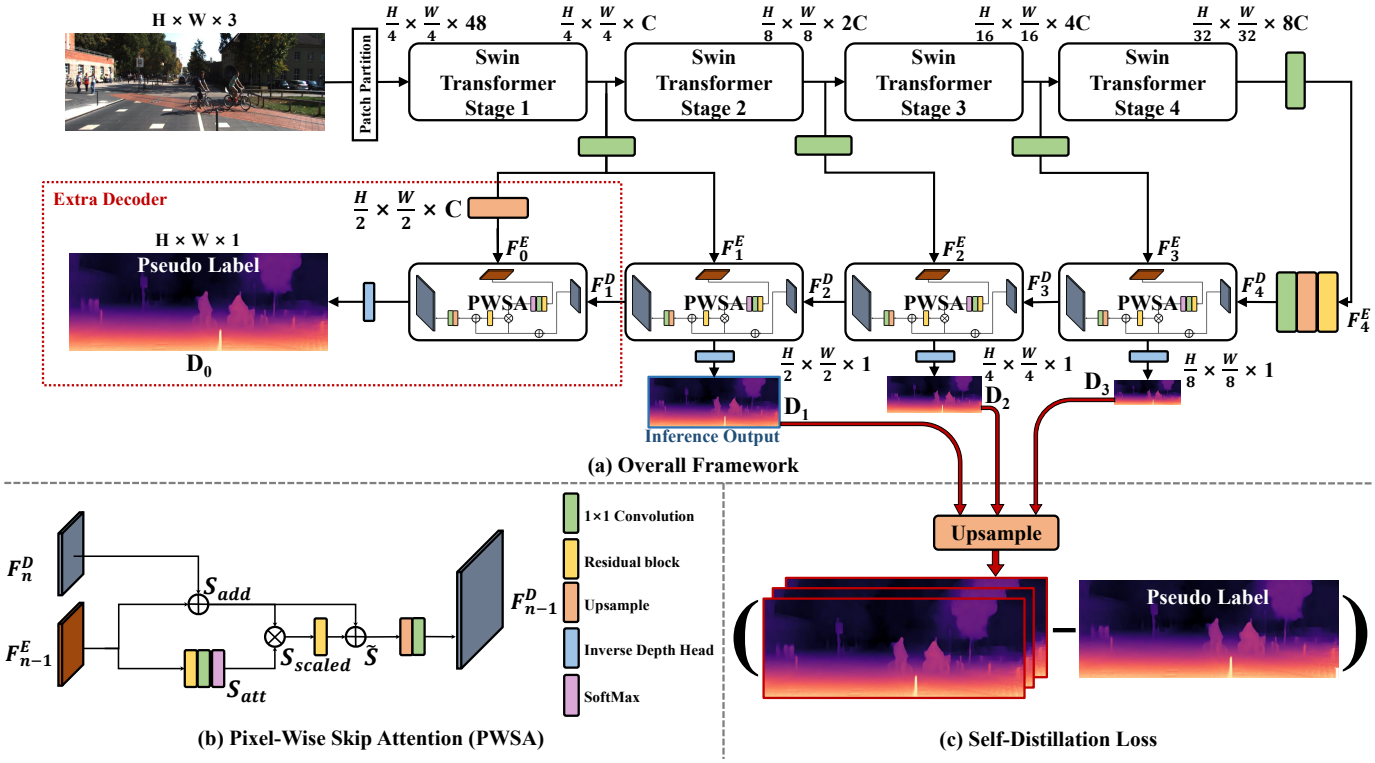


Fig. 2. **The overview of the proposed framework**. **(a) Overall Framework**: The proposed model consists of 4 sets of a swin transformer encoder that is a hierarchical structure and the proposed PWSA module as the decoder. **(b) Pixel-Wise Skip Attention (PWSA)**: The PWSA module takes a pair of encoder/decoder features with the same resolution, and encourages the model to learn fine-grained information by extracting the pixel-wise attention map. **(c) Self-distillation loss**: For mitigating the high complexity of the model in the inference step and emphasizing fine details in multi-scale inverse depth predictions, we distill the finer details of the highest-resolution inverse depth image into the multi-scale inverse depth images.

B. Model Architecture

Overall Architecture: Adopt Transformer We propose a transformer-based architecture designed for a monocular depth estimation task to effectively aggregate global and local contexts. For global context, we employ Swin Transformer [10] as our backbone encoder, which is efficient due to the window-based self-attention and favorable to learning various scales of the visual entity from its hierarchical feature maps.

As shown in Fig. 2-(a), a given input image is fed into the network. Then, multi-scale encoder feature maps F_n^E are generated after every encoder stage followed by a 1×1 convolution layer. Since the resolution of each feature map is downsampled by a factor of 2 from the first feature map, as in conventional CNN-based encoders, this Swin Transformer encoder can easily replace convolution-based encoders. After passing through the encoders, the last encoder feature map $F_4^E \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32}}=$ is up-sampled after a residual block, then another 1×1 convolution is applied to form a decoder feature map $F_4^D \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16}}$. An encoder feature F_{n-1}^E and decoder feature F_n^D pair is used as input of a decoder module which produces a decoder feature map F_{n-1}^D. This output feature map F_{n-1}^D is used as an input of both the next decoder stage and an inverse depth head.

Pixel-Wise Skip Attention Adopting transformers is not sufficient for the dense depth estimation task because of the lack of local details, as shown in Fig. 1. Since our goal is to make the decoder feature map F_n^D contain rich global and local contexts, we propose Pixel-Wise Skip Attention (PWSA), a simple yet effective attention block, and use it as a decoder module to enhance local details.

Our PWSA module, as illustrated in Fig. 2-(b), consists of two streams; long skip connection and pixel-wise attention. The long skip connection is implemented by element-wise summation S_{add} of two input feature maps F_n^D and F_{n-1}^E. Another stream is to calculate pixel-wise spatial attention map S_{att}, which is known to be good at encoding location-aware information in general, by applying a residual block and linear-embedding followed by softmax to F_{n-1}^E. This attention map S_{att} helps the summed feature S_{add} convert into a spatially detail enhanced feature S_{scaled}.

$$\begin{aligned} S_{add} &= F_{n-1}^E + F_n^D \\ S_{att} &= (softmax \circ embedding \circ resblock)(F_{n-1}^E) \\ S_{scaled} &= S_{add} \odot S_{att} \end{aligned} \tag{2}$$

where $resblock$ consists of two 3×3 convolution layers each followed by ReLU activation, $embedding$ is implemented by a 1×1 convolution layer, and $\odot$ indicates element-wise multiplication (Hadamard product).

To mitigate the rescaled feature S_{scaled} from being overly emphasized or ignored at the early stage of training reported in [21], we apply another $resblock$ to S_{scaled} and add a fused feature before attention S_{add} to that. Finally, the output decoder feature map F_{n-1}^D is obtained by applying bilinear up-sampling by 2 followed by another 1×1 convolution to the

enhanced feature map $\tilde{S}$ as follow:

$$\begin{aligned} \tilde{S} &= resblock(S_{scaled}) + S_{add} \\ F^D_{n-1} &= (embedding \circ upsample_2)(\tilde{S}) \end{aligned} \tag{3}$$

We visualize feature maps with and without the PWSA module in Fig. 3. The feature maps without PWSA in the second row are activated globally, and the boundary of objects is not clearly visible. On the other hand, object boundaries in feature maps with PWSA shown in the third row look distinguishable from background pixels. In short, our PWSA helps improving fine details in feature maps.

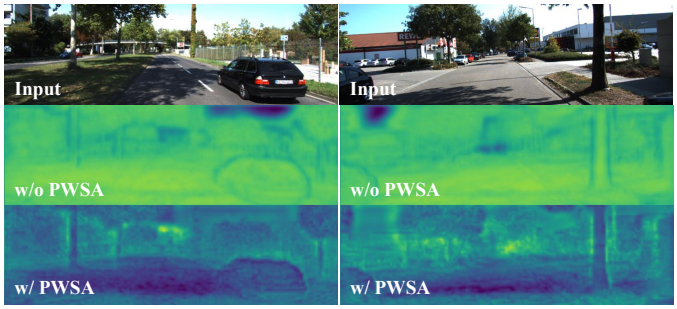


Fig. 3. **Visualization of feature maps with and without PWSA module.** The feature without PWSA is globally activated, while the feature map with PWSA shows that the local detail regions are activated.

C. Self-Distillation Loss with Pseudo Label

The photometric loss is known as a weak training signal [13], *i.e*, optimizing the depth network by photometric error does not always lead to predicting the correct depth perfectly. For example, textureless/occluded regions and moving objects hinder the model from being trained correctly, even though auto-mask [3] can alleviate learning from them. Our preliminary experiment, shown in Fig. 1 and first row in Table III, observed that multi-scale photometric loss results in inferior performance for a naive transformer architecture, *i.e*, "transformer-baseline". Due to the transformer's instability in training [24], [36], we conjecture that the weak training signal from photometric loss is unfavorable for transformers to learn accurate intermediate feature representations, e.g., F^E_n and F^D_n.

We propose a novel self-distillation loss that uses the last inverse depth prediction, *i.e*, the highest resolution inverse depth map, as a pseudo label. Our basic idea is that the prediction in the largest resolution, which greatly contributes to complexity, usually shows the best performance, and we can transfer the capability of predicting fine details in the highest resolution inverse depth to the penultimate prediction by distillation. We propose to replace the multi-scale photometric loss with a combination of single-scale photometric loss and self-distillation loss. The prediction with the highest resolution is still supervised by single-scale photometric loss and used as a pseudo label to supervise other intermediate predictions with different resolutions. It prevents learning from the repeated weak training signals caused by photometric loss. Our self-distillation loss shows better depth estimation performances by learning better intermediate representations and helps improve the instability of transformer training.

To calculate the self-distillation loss $\mathcal{L}_{sd}$, we upsample inverse depth predictions $D_i,\ i \in [1,2,3]$ except for the highest resolution of prediction D_0 to match the resolution with the pseudo label D_0. This is to keep as many fine details in the pseudo label as possible. L_1 loss is then calculated between the upsampled predictions and the pseudo label. For training stability, the auto-mask M [3] is employed. In addition, we propose an adaptive weighting scheme λ_{adp} which controls the loss weights gradually over the training progress to alleviate the undesired error propagation due to the inaccuracy of pseudo label at the early stage of training as follow:

$$\begin{aligned} \mathcal{L}_{sd} &= \lambda_{adp} * \frac{1}{3}\sum_{i=1}^{3} |upsample(D_i, 2^i) - D_0|_1 * M \\ \lambda_{adp} &= \begin{cases} epoch/epoch_{thr}, & epoch < epoch_{thr} \\ 1, & othersize \end{cases} \end{aligned} \tag{4}$$

where, we set $epoch_{thr}$ to 10.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

KITTI The KITTI benchmark [11] provides a large-scale outdoor multi-sensor dataset that can be used for monocular depth estimation. We follow the standard protocol to decide the training/testing split proposed by Eigen *et al.* [2] as well as Zhou *et al.* [12]'s pre-processing to remove static frames. Thus we use $39,810$ frames for training and 697 frames for evaluation. We evaluated the models in the 0-80m range.

DDAD The DDAD dataset contains diverse scenarios such as urban, highway, and residential areas collected by fleets in the US and Japan. This dataset consists of 17,050 frames for training and 4,150 frames for evaluation. Since it is designed for long-range monocular depth estimation, performances are evaluated in the 0-200m range.

Evaluation Metrics For evaluation, we follow the standard evaluation metrics proposed by Eigen *et al.* [2]. Please note that lower results are better in error metrics such as AbsRel, SqrRel, RMSE, and RMSE log, but higher results are better in recall metrics such as $\delta_{1.25}$ (denoting $\delta < 1.25$).

B. Implementation Details

We used the same pose network proposed in [12] without the explainability mask. Each inverse depth head consists of two 3×3 convolution layers followed by a disparity discrete volume [1]. The standard appearance matching loss [5], [12] was used, which consists of a structural similarity index measurement (SSIM) loss and L1 loss with edge-aware depth smoothness regularization [5]. We utilized per-pixel minimum reprojection loss [3] to remove occluded or out-of-view pixels, and employed the velocity loss [6] to improve training stability by supervising absolute translation in the predicted pose.

TABLE I
QUANTITATIVE RESULTS ON KITTI EIGEN SPLIT SWIN-LARGE-22K IS INITIALIZED BY PRETRAINED WEIGHTS ON IMAGENET22K DATASET. LOWER IS BETTER FOR METRICS IN RED, HIGHER IS BETTER FOR METRICS IN BLUE. THE BEST AND THE SECOND-BEST ARE HIGHLIGHTED. ALL THE METHODS ARE TRAINED FROM MONOCULAR VIDEOS WITH 640×192 RESOLUTION BY SELF-SUPERVISION.

Method	Arch.Change	Transformer	Abs. Rel ↓	Sqr. Rel ↓	RMSE ↓	RMSE log ↓	$\delta_{1.25}$↑	$\delta_{1.25^2}$↑	$\delta_{1.25^3}$↑
Monodepth2 [3]	-	-	0.115	0.903	4.863	0.193	0.877	0.959	0.981
RSDE [19]	-	-	0.108	0.737	4.562	0.187	0.883	0.961	0.982
PackNet-SfM [6]	✓	-	0.111	0.785	4.601	0.189	0.878	0.960	0.982
ADAADepth [17]	✓	-	0.111	0.817	4.685	0.188	0.883	0.961	0.982
MLDA-Net [4]	✓	-	0.110	0.824	4.630	0.187	0.883	0.961	0.982
HR-Depth [34]	✓	-	0.109	0.792	4.632	0.185	0.884	0.962	0.983
DDV [1]	✓	-	0.106	0.861	4.699	0.185	0.889	0.962	0.982
CADepth-Net [23]	✓	-	0.105	0.769	4.535	0.181	0.892	0.964	0.983
DIFFNet [26]	✓	-	0.102	0.764	4.483	0.180	0.896	0.965	0.983
MT-SfMLearner [37]	✓	✓	0.112	0.838	4.771	0.188	0.879	0.960	0.982
Ours (Swin-Tiny)	✓	✓	0.102	0.753	4.461	0.177	0.896	0.966	**0.984**
Ours (Swin-Small)	✓	✓	0.098	0.728	4.458	0.176	0.898	0.966	**0.984**
Ours (Swin-Large-22k)	✓	✓	**0.095**	**0.711**	**4.321**	**0.172**	**0.906**	**0.967**	**0.984**

TABLE II
QUANTITATIVE RESULTS ON DDAD BENCHMARK. ALL THE METHODS ARE TRAINED FROM MONOCULAR VIDEOS WITH 640×384 RESOLUTION BY SELF-SUPERVISION.

Method	Abs. Rel ↓	Sqr. Rel ↓	RMSE ↓	RMSE log ↓
Monodepth2 [3]	0.227	11.293	17.368	0.303
PackNet-SfM [6]	0.173	4.164	14.363	0.249
Transformer-baseline	0.165	3.995	14.758	0.252
Ours (Swin-S)	**0.151**	**3.591**	**14.350**	**0.244**

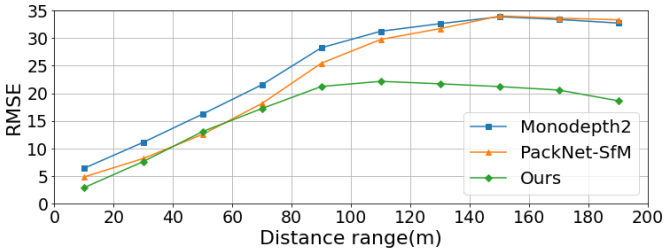


Fig. 4. **Depth prediction accuracy over various distance ranges in DDAD.** Each marker represents the average of the RMSE at 20-meter intervals, e.g., 0-20m, 20-40m, and so on.

We used PyTorch to implement our models, and the models were trained on a single NVIDIA RTX 3090 GPU. We used an AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. All the models were trained for 22 epochs with batch size 4. The learning rates for the pose network, transformer encoder, and PWSA decoder set to 10^{-4}, 10^{-5}, and 6×10^{-5}, respectively, and decayed by half at 18 epochs. The stride for training video clips sets to 1, which indicates that the previous $t-1$, current t, and next $t+1$ images in time are used as a single data point for training. We set the depth regularization weight to 0.001 and the velocity scaling weight to 0.1.

C. Depth Estimation Performance

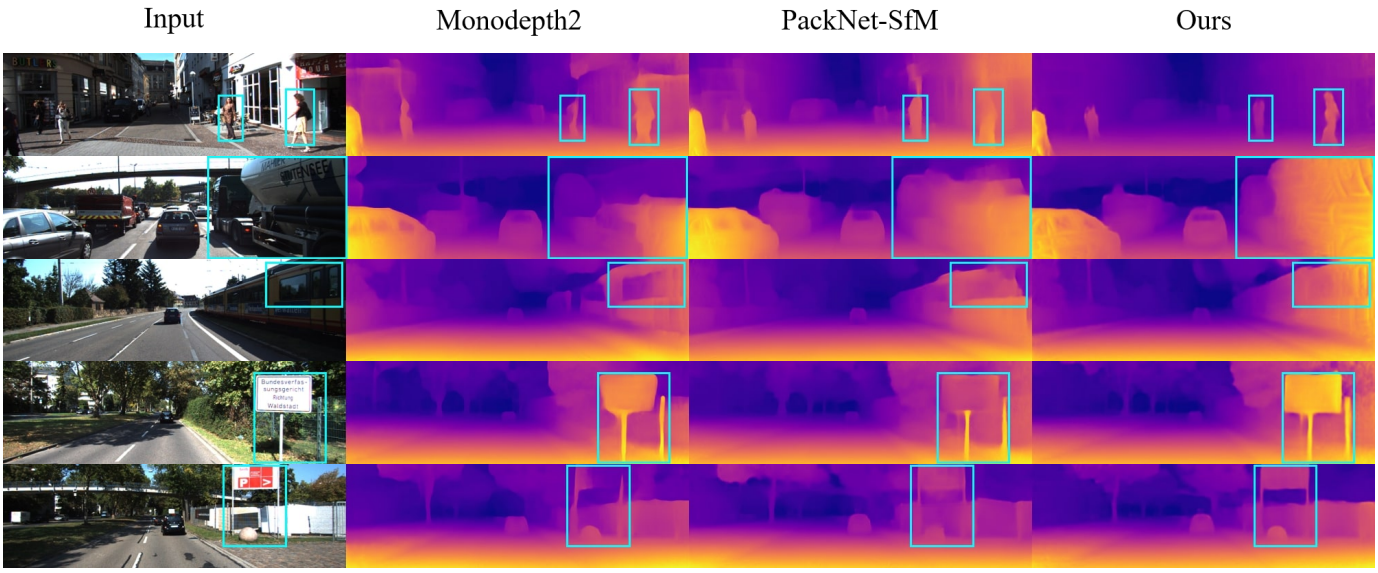
KITTI Eigen Split First, we train and evaluate our models on the KITTI dataset. As shown in Table I, our models achieve state-of-the-art performance in all metrics compared to previous CNN-based self-supervised monocular depth estimation methods [3], [19], including architectural changes such as attention modules [1], [4], [17], [23], [26]. Our models especially outperform the previous transformer-based method [37] by a large margin, which does not carefully consider transformer adaption, such as ours. Our model achieves a new state-of-the-art performance in the self-supervised monocular depth estimation task.

In Fig. 5-(a), we present qualitative results on the KITTI dataset. Since our model has a strong ability to aggregate global context effectively from the transformer-based encoder while keeping the local details due to our PWSA decoder modules, our model clearly sees the shape of the pedestrians on the right in the first row. In addition, our model successfully predicts the correct shape of the oil tanker and trains in the second and third rows, thanks to the transformer's global context encoding capability. Furthermore, as shown in the fourth and last case, our model also makes a good prediction on thin vertical structures, which requires enhanced local details. We show that our model, which has transformer encoders and PWSA decoders trained by single-scale photometric loss with self-distillation loss, has an advantage of estimating an accurate dense depth map. Ablation studies are discussed in the subsection, Sec. IV-D.

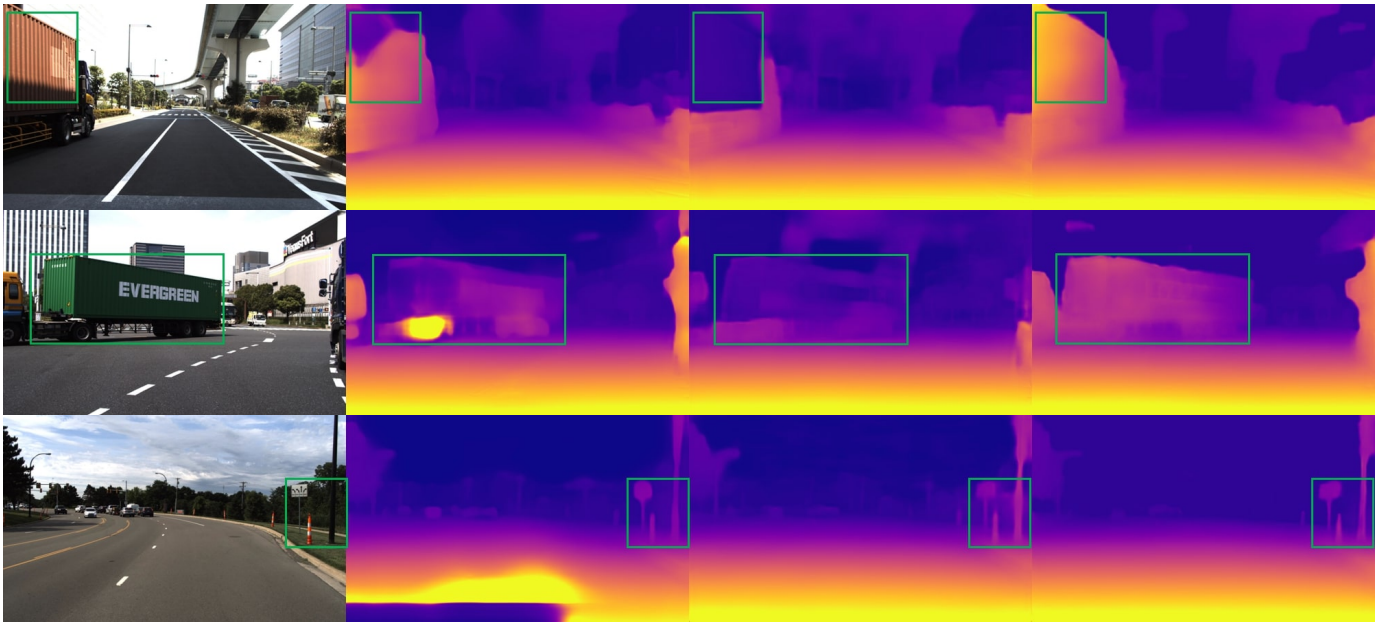
DDAD We also report the performance on the DDAD dataset. Our model significantly outperforms existing methods, as shown in Table II, including an advanced CNN-based network, PackNet-SfM [6], by a large margin (12.7% improvement, 0.173 → 0.151 in Abs Rel).

In Fig. 4, we measure the average RMSE values for each distance range, e.g., 0-20m, 20-40m. Our model outperforms the previous models consistently over all distance ranges, except for the 40-60m range. Considering that the only way to predict depth on very far-away regions (>100m) is to utilize the global context of the scene, this shows that our transformer-based model is better at to capturing global context.

In Fig. 5-(b), we also illustrate qualitative depth predictions on DDAD. Compared to the KITTI dataset, the DDAD is taken in more challenging and realistic conditions, so that previous



(a) KITTI Eigen split



(b) DDAD

Fig. 5. **Qualitative results.** The proposed model produces better predictions on large objects (e.g., the oil tank, train, container) that require seeing the global context and thin vertical structures (e.g., pedestrian, traffic sign, pole) that need fine-detailed features than previous methods.

self-supervised learning approaches did not work well. On the other hand, we can see the same improvement on large objects, such as the containers in the first and second rows, by effectively capturing the larger context. In addition, as shown in the bottom row, our model shows reasonable results in the global context (e.g. road) and local details (e.g. sign), while prior methods produce noisy depth results.

D. Ablation Study

In our preliminary experiment shown in Fig. 1 and Table III (2nd row), we replace convolution encoders to the transformer as a naive approach, called "transformer baseline", to benefit from using a transformer. However, the performance gain is marginal (0.110 $\rightarrow$ 0.108 in AbsRel) or worse (0.840 $\rightarrow$ 1.004 in SqrRel, 4.765 $\rightarrow$ 4.838 in RMSE), as shown in Table III. We argue that this result is due to 1) the limited ability to capture local details in a typical decoder module and 2) applying photometric loss function in a multi-scale manner. We propose the use of single-scale photometric loss rather than multi-scale loss, PWSA as an enhanced decoder, and self-distillation loss to replace weak supervision by photometric loss. We conducted an ablation study on the proposed component as follows.

Multi-scale loss vs. Single-scale loss As discussed in Sec. III-C, the typical photometric loss is a weak training

TABLE III
ABLATION STUDY OF THE PROPOSED METHOD ON KITTI. IF "DETACH" IS MARKED, PERFORMANCE IS EVALUATED FROM THE PENULTIMATE PREDICTION (D_1), NOT FROM THE LAST PREDICTION (D_0). THE BASELINE MODEL IS HIGHLIGHTED.

(Rel.%)

SS Loss	PWSA	SD Loss	Detach	Backbone (Params)	Abs. Rel ↓	Sqr. Rel ↓	RMSE ↓
Baseline [1]				DR101(64M)	0.110 (±0)	0.840 (±0)	4.765 (±0)
-	-	-	-	Swin-S(55M)	0.108 (↓ 1.8)	1.004 (↑ 19.5)	4.838 (↑ 1.5)
✓	-	-	-	Swin-S(55M)	0.104 (↓ 5.5)	0.882 (↑ 5.0)	4.639 (↓ 2.6)
✓	✓	-	-	Swin-S(60M)	0.103 (↓ 6.4)	0.804 (↓ 4.3)	4.558 (↓ 4.3)
✓	✓	✓	-	Swin-S(62M)	0.099 (↓ 10.0)	0.758 (↓ 9.8)	4.483 (↓ 5.9)
✓	✓	✓	✓	Swin-S(60M)	**0.098** (↓ 10.9)	**0.728** (↓ 13.3)	**4.458** (↓ 6.4)

SS Loss: Single-scale photometric loss
SD Loss: Self-Distillation loss
Baseline: DDV without self-attention [1]
DR101: ResNet101 with dilated convolution

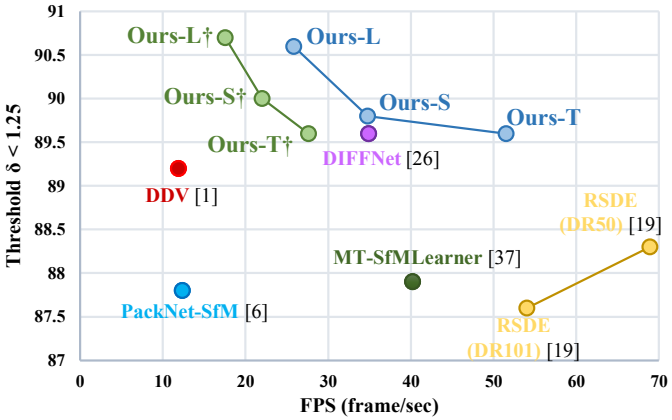


Fig. 6. **Performance of depth networks for varying runtime on KITTI Eigen split [2].** (DR50 and DR101) indicate (Deformable ResNet50 and Deformable ResNet101) [19] respectively. † means pseudo label (*i.e*the highest resolution prediction D_0) is used for evaluation without detaching the last decoder. The threshold $\delta < 1.25$ is detailed in [2]: higher is better.

signal, which has a clear limitation. In addition, since the photometric loss is usually applied in a multi-scale manner to help intermediate features capture better information, it might disturb the intermediate features from learning a good representation from such a weak training signal. As an alternative to the multi-scale loss, we apply single-scale photometric loss only to the highest resolution of the prediction (D_0 in Fig. 2) to ease the disturbing effect from weak-supervision. Surprisingly, as shown in Table III, the single-scale photometric loss shows better performance (3nd row, 0.104 Abs.Rel) than the multi-scale baseline (2nd row, 0.108). This supports our hypothesis and we use the single-scale loss for the remainder of the experiments.

Effect of PWSA decoder On top of the single-scale loss, we achieve greater performance gain by employing the proposed PWSA decoder (4th row, 0.103 in Abs.Rel). Compared to the model without PWSA decoder (3rd row, 0.104 in Abs.Rel), this model shows significant improvement in another metric (8.8%, 0.882 → 0.804 in Sqr. Rel). This result indicates that some of the high-error predictions are improved; in other words, the PWSA module helps predict spatially coherent depth. Since most of the high-error points are on the object boundaries, this indicates a quantitative improvement by sharp depth discontinuity, as shown in Fig. 1.

Advantage of Self-Distillation Loss We decided to use the single-scale loss from the above experiments for better performance. Then there is no supervision for the intermediate feature maps. In this case, our self-distillation loss could be helpful, and we employ the adaptive weighting scheme, which increases gradually over the training progress to avoid error propagation due to premature pseudo labels at the early stage of training. Since our self-distillation loss can supervise the intermediate predictions to have the correct depth using the pseudo label regardless of textures such as walls or ground plane, another improvement is achieved by applying the self-distillation loss shown in Table III (5th row, 0.103 → 0.099 in Abs.Rel, 0.804 → 0.758 in Sqr.Rel, and 4.558 → 4.483 in RMSE).

Another interesting advantage of applying self-distillation loss is, as discussed in Sec. III-C, the possibility to transfer the capability of predicting fine details to the penultimate output. We could expect both predictions, the highest resolution output and the penultimate resolution output, to show similar performance after training. If so, we can detach the last decoder at inference time and use the penultimate resolution output as our final prediction. Since this detachable property is a by-product of self-distillation, we could treat the last decoder as an "extra" decoder which is too heavy to deploy but worth adding during training. Throughout this experiment, we assume the last decoder to produce D_0 is an extra decoder, requiring an additional 2M parameters in the 5th row. We compare performances of the highest resolution (D_0 in Fig. 2-(a), 5th row in Table III) and the penultimate resolution (D_1 in Fig. 2-(a), 6th row in Table III). Surprisingly, the penultimate output performs better with fewer parameters (62M vs. 60M). This supports our conjecture that our self-distillation loss as an alternative to the photometric loss could be a better training signal that always guides to the correct depth regardless of the scene contents, e.g., textures.

Speed vs. Accuracy Trade-off The speed-accuracy trade-off is an the important aspects of practicality. In Fig. 6, we mark various self-supervised monocular depth estimation models in an accuracy versus inference time plot on the KITTI benchmark. For a fair comparison, all the models are evaluated on the same NVIDIA RTX 2080Ti GPU. Since our models, trained by the self-distillation loss, have the "detachable extra decoder", we evaluate both predictions with and without the last decoder. As shown by the blue circles in Fig. 6, our models run in real-time (>20fps) while keeping accuracy when detaching the extra decoder. Also, our models outperform other methods in accuracy and show competitive efficiency.

V. CONCLUSIONS

Here, we propose an effective way to employ transformer architecture in self-supervised monocular depth estimation. The proposed architecture effectively leverages global context from transformer-based encoders and fine details from our

PWSA decoder. We propose to use self-distillation loss with a single-scale photometric loss instead of a typical multi-scale photometric loss. In our experiments, our final model shows improved depth predictions on the far range and the object boundaries and outperforms existing self-supervised monocular depth estimation methods in terms of accuracy with competitive efficiency. We expect our method to be a cornerstone of transformer-based self-supervised monocular depth estimations.

REFERENCES

[1] A. Johnston and G. Carneiro, "Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[2] D. Eigen, C. Puhrsch and F. Rob, "Depth map prediction from a single image using a multi-scale deep network," in *Proceeding of Conference on Neural Information Processing Systems*, 2014.

[3] C. Godard, O. Aodha, M. Firman and G. Brostow, "Digging Into Self-Supervised Monocular Depth Estimation," in *Proceeding of IEEE/CVF International Conference on Computer Vision*, 2019.

[4] X. Song, W. Li, D. Zhou, Y. Dai, J. Fang, H. Li and L. Zhang, "MLDA-Net: Multi-Level Dual Attention-Based Network for Self-Supervised Monocular Depth Estimation," in *IEEE Transactions on Image Processing*, vol. 30, pp. 4691-4705, 2021.

[5] C. Godard, O. Aodha and G. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceeding of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[6] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos and A. Gaidon, "3D Packing for Self-Supervised Monocular Depth Estimation," in *Proceeding of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[7] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834-848, 2018.

[8] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, "Pyramid scene parsing network," in *Proceeding of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[9] R. Ranftl, B. Alexey and V. Koltun, "Digging Into Self-Supervised Monocular Depth Estimation," in *Proceeding of IEEE/CVF International Conference on Computer Vision*, 2021.

[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proceeding of IEEE/CVF International Conference on Computer Vision*, 2019.

[11] A. Geiger, P. Lenz, C. Stiller and R. Urtasun, "Vision meets robotics: The kitti dataset," in *International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231-1237, 2013.

[12] T. Zhou, M. Brown, N. Snavely and D. Lowe, "Unsupervised learning of depth and egomotion from video," in *Proceeding of IEEE Conference/CVF on Computer Vision and Pattern Recognition*, 2017.

[13] C. Shu, K. Yu, Z. Duan and K. Yang, "Feature-metric loss for self-supervised learning of depth and egomotion," in *Proceeding of European Conference on Computer Vision* , 2020.

[14] H. Chawla, A. Varma, E. Arani and B. Zonooz, "Multimodal Scale Consistency and Awareness for Monocular Self-Supervised Depth Estimation," in *Proceeding of IEEE International Conference on Robotics and Automation*, 2020.

[15] J. Hur and S. Roth, "Self-Supervised Monocular Scene Flow Estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[16] V. Guizilini, R. Hou, J. Li, R. Ambrus and A. Gaidon, "Semantically-guided representation learning for self-supervised monocular depth," in *Proceeding of International Conference on Learning Representations*, 2020.

[17] V. Kaushik, K. Jindgar and B. Lall, "ADAADepth: Adapting Data Augmentation and Attention for Self-Supervised Monocular Depth Estimation," in *IEEE Robotics and Automation Letters*, vol. 6, pp. 7791-7798, 2021.

[18] G. Yang, H. Tang, M. Ding, N. Sebe, and E. Ricci, "Transformer-based attention networks for continuous pixel-wise prediction," in *Proceeding of IEEE/CVF International Conference on Computer Vision* , 2021.

[19] U. Kim and J. Kim, "Revisiting Self-Supervised Monocular Depth Estimation," in *arXiv preprint arXiv:2103.12496*, 2021.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proceeding of International Conference on Learning Representations*, 2021.

[21] S. Huang, Z. Lu, R. Cheng and C. He, "FaPN: Feature-aligned Pyramid Network for Dense Image Prediction," in *Proceeding of IEEE/CVF International Conference on Computer Vision*, 2021.

[22] R. Guo, D. Niu, L. Qu, and Z. Li "Sotr: Segmenting objects with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[23] J. Yan, H. Zhao, P. Bu and Y. Jin, "Channel-Wise Attention-Based Network for Self-Supervised Monocular Depth Estimation," in *Proceeding of International Conference on 3D Vision*, 2021.

[24] T. Xiao, P. Dollar, M. Singh, E. Mintun, T. Darrell and R. Girshick "Early convolutions help transformers see better," in *Advances in Neural Information Processing Systems*, 2021.

[25] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei and C. Shen "Twins: Revisiting the design of spatial attention in vision transformers, " in *Advances in Neural Information Processing Systems*, 2021.

[26] H. Zhou, D. Greenwood, and S. Taylor, "Self-Supervised Monocular Depth Estimation with Internal Feature Fusion," in *Proceeding of British Machine Vision Conference*, 2021.

[27] X. Mao, G. Qi, Y. Chen, X. Li, R. Duan, S. Ye and H. Xue "Towards robust vision transformer," *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2022.

[28] X. Ma, Z. Wang, H. Li, P. Zhang, W. Ouyang and X. Fan "Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving" *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[29] W. Wang, E. Xie, X. Li, D. P. Fan, K. Song, D. Liang, K. Song, D. Liang, T. Lu, P. Luo and L. Shao "PVT v2: Improved baselines with Pyramid Vision Transformer" *Computational Visual Media*, 2022.

[30] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. FU, J. Feng, T. Xiang, P. Torr and L. Zhang "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers" *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser and I. Polosukhin, I. "Attention is all you need" *Advances in neural information processing systems*, 2017.

[32] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez and P. Luo "SegFormer: Simple and efficient design for semantic segmentation with transformers" *Advances in Neural Information Processing Systems*, 2021.

[33] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen and J. Wang "HRFormer: High-Resolution Vision Transformer for Dense Predict" *Advances in Neural Information Processing Systems*, 2021

[34] 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" *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021

[35] Y. Wang, W. Chao, D. Garg, B. Hariharan, M. Campbell and K. Q. Weinberger "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *Proceeding of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[36] L. Liu, X. Liu, J. Gao, W. Chen and J. Han "Understanding the Difficulty of Training Transformers," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2020.

[37] A. Varma, H. Chawla, B. Zonooz, and E. Arani,sd "Transformers in Self-Supervised Monocular Depth Estimation with Unknown Camera Intrinsics" *In Proceedings of the International Conference on Computer Vision Theory and Applications*, 2022.

[38] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu "Incorporating convolution designs into visual transformers." in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.