Abstract—Depth estimation using monocular camera sensors is an important technique in computer vision. Supervised monocular depth estimation requires a lot of data acquired from depth sensors. However, acquiring depth data is an expensive task. We sometimes cannot acquire data due to the limitations of the sensor. View synthesis-based depth estimation research is a self-supervised learning method that does not require depth data supervision. Previous studies mainly use the convolutional neural network (CNN)-based networks in encoders. The CNN is suitable for extracting local features through convolution operation. Recent vision transformers (ViTs) are suitable for global feature extraction based on multiself-attention modules. In this article, we propose a hybrid network combining the CNN and ViT networks in self-supervised learning-based monocular depth estimation. We design an encoder–decoder structure that uses CNNs in the earlier stage of extracting local features and a ViT in the later stages of extracting global features. We evaluate the proposed network through various experiments based on the Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) and Cityscapes datasets. The results showed higher performance than previous studies and reduced parameters and computations. Codes and trained models are available at https://github.com/fogfog2/manydepthformer.

Index Terms—Depth estimation, monocular sensor estimation, self-attention, self-supervised, transformer.

I. INTRODUCTION

Depth estimation is an important technology used for various fields that require 3-D spatial map generation, such as autonomous driving, augmented reality (AR), virtual reality (VR), and robot vision. The conventional sensor-based depth estimation method has high accuracy but has problems of high price and low resolution. Recently, camera-based depth estimation research has been actively conducted, and the performance is also high enough to be applied to commercial products such as drones. The traditional stereo camera-based method estimates the depth by calculating the disparity between two calibrated cameras. A recent monocular camera-based method estimates disparity between temporally adjacent frames. Some monocular camera-based depth estimation algorithms are trained with self-supervision to avoid difficulties in acquiring depth data.

The monocular camera-based depth estimation study is divided into a supervised learning-based method and an unsupervised learning-based method. Supervised learning-based networks directly learn depth data obtained from sensors. The unsupervised learning-based method estimates depth based on a viewpoint synthesis technique that reconstructs the current image from temporally adjacent consecutive frames [1]. The supervised learning-based method outputs higher performance than unsupervised learning because it learns directly from depth data. However, acquiring 3-D data is an expensive operation. When 3-D data cannot be acquired, or a depth sensor is not available, unsupervised learning is the only solution. Recently, research on unsupervised learning-based models is being actively conducted in various fields such as automobile autonomous driving [2], the smartphone-based AR system [3], the drone avoidance system [4], and the medical system [5]. However, there is a problem in that the unsupervised learning method based on the monocular camera cannot estimate the absolute distance of the depth. Therefore, the output of unsupervised learning depth estimation is limitedly applicable to fields that are available with relative depth. Some studies additionally train velocity to estimate absolute depth or use inertial measurement unit (IMU) sensors [6].

In most cases, the convolutional neural network (CNN)-based backbone was the basic model of the existing deep-learning-based monocular camera depth estimation. This was true not only in the field of depth estimation [7], but also in the field of image classification [8], segmentation [9], and detection [10]. Two CNN-based residual network (ResNet)
[11] and EfficientNet [12] are widely used as the main backbone models for deep-learning networks. Convolution operation represents global features map in a structure that increases the receptive area by overlapping deep layers based on the association of local pixels. This structure has traditionally been a general structure with state-of-the-art performance.

Transformer is a recently popular backbone network in the field of depth estimation [13], [14]. The original transformer is a model proposed in the field of natural language processing to predict sequences. Some transformer models showed high performance in the natural language processing field, and the model applied to the computer vision field is the vision transformer (ViT) [15]. The ViT applied the same format used for natural language processing to the vision field. It consists of a multilayer attention mechanism and a multilayer perceptron (MLP). In ViT, multilayer attention extracts global features, and MLP expresses local features just like CNNs. A recent ViT model has achieved the highest performance in image classification, segmentation, and detection [16], [17]. However, the ViT outperforms the CNN-based backbone when there is a lot of training data, and it is necessary to solve the problem of the high computational amount of the multilayer attention mechanism.

In an early study applying a ViT to supervised learning of monocular camera depth estimation, a hybrid model that mixed CNN-based ResNet and ViT models showed the best performance by [14]. A recent study improved the depth estimation performance by using a Swin transformer applied with a hierarchical encoder–decoder and multiscale fusion attention [13]. In this article, following the above studies, we design an efficient self-supervised depth estimation network using lightweight attention (LA). Additionally, the proposed network shows higher performance by using the expert layer for each encoder layer depth.

In this article, we propose a hybrid transformer-based depth estimation network for effective depth network estimation in an unsupervised depth estimation model. First, the proposed network consists of a cost-volume structure for depth and a pose network for estimating poses between temporally adjacent images. Next, the feature extractor transforms input images into feature maps. The feature map of the source image is synthesized with the viewpoint of the target image. In this feature map view synthesis process, the predicted pose, assumed depth, and camera parameters are used. The cost volume is generated by accumulating the difference between the source feature maps synthesized for each depth and the target feature map. The target feature map and cost volume are concatenated. The combined features are transformed into depth in the proposed hybrid encoder–decoder. Encoders are connected to layers of networks with different properties. In the early layer of the encoder, a CNN-based ResNet backbone is used for local representation. The later layers use transformers for global representation. We apply a lightweight transformer model to solve the problem of the high computational complexity of multilayer attention mechanisms. The lightweight transformer reduces the number of parameters and operations compared to the ResNet layer. We can design a more efficient network by applying a lightweight model. In the decoder, channel-space attention is additionally applied to the commonly used multiscale confusion module to improve the depth estimation performance.

The structure of this article is as follows. Section II investigates related studies, backbone networks, and unsupervised depth estimation. Section III describes the structure of the proposed overall network, and Section IV shows the experimental results of applying the proposed method to the Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) and Cityscapes datasets. The conclusion of the proposed method is mentioned in Section V.

II. RELATED WORK

A. CNN and Transformer

Since the proposal of AlexNet [18], the CNN has been used as the main backbone network in the field of computer vision. Various models have been studied, such as visual geometry group (VGG) [19], ResNet [11], MobileNet [20], and EfficientNet [12]. VGG analyzed the effect of the depth of the network, and ResNet proposed a residual network that merges the input with the output. MobileNet improves network efficiency with depthwise convolution and an inverted residual block. EfficientNet improves performance by compound scaling that determines the depth, width, and size of the input image of the network. Recent studies use the above general CNN-based backbone networks with some modifications [7], [8]. An ensemble model is also used for optimization [21].

In particular, ResNet improves the learning speed and training effect of the network without significantly increasing parameters and computations with a shortcut structure [11], [22]. Also, even when the number of model layers increases, gradient vanishing is prevented by the effect of skip connection. In the field of depth estimation, the ResNet backbone is mainly used following the main contribution paper Monodepth2 [23]. Following the previous works, we construct an efficient hybrid network based on ResNet.

Transformer is a model that showed good performance in the field of natural language processing. ViT is an early ViT model that showed the best performance by applying a transformer in the image classification field. Data-efficient image transformers (DeiT) [24] proposed a method for knowledge distillation of the CNN-based network results on a transformer and applied various data augmentation techniques. Swin transformer proposes a hierarchical feature map that reduces image resolution by patch mixing for each stage. With the proposal of the hierarchical feature map, it became possible to apply the transformer not only to the classification field, but also to the image detection and segmentation field [16]. In addition, the Swin transformer proposes a cyclic shifting window to improve local feature expression performance. After that, convolutions to vision transformers (CvT) [25] removes the limitation of nonoverlapping patch unit embeddings by using convolution for token embedding. A recent study, CNNs meet Vision transformers (CMT) [26], proposed lit multilayer self-attention that reduces the spatial resolution of keys and values. In addition, an inverted residual feed-forward network (IRFFN) based on depthwise separable convolution is used instead of a feed-forward network to improve local representation.
B. Self-Supervised Monocular Depth Estimation

Although the supervised learning method shows relatively good performance in the field of monocular camera depth estimation, in a recent study, the unsupervised learning method also shows comparable performance [27]. The unsupervised learning model is a depth estimation method that can be easily applied to images for which it is not easy to acquire depth data.

Garg et al. [28] proposed a viewpoint synthesis technique for unsupervised learning-based depth estimation in stereo images. This method requires stereo pair images for training, but the network estimates the depth in a single frame at test time. A reconstructed left image is generated using the geometric constraint between the right image and the left depth image estimated from the depth network. The reconstruction error between the reconstructed left image and the left image is used as a loss function of the depth network. Godard et al. [29] applied the spatial transformer network (STN) [30] to sampling as an image reconstruction method, and they also proposed a photometric loss combining structural similarity index measure (SSIM) [31] and L1 loss. Tosi et al. [32] and Watson et al. [33] proposed a semi-global matching-based depth estimation method that can generate different depths according to hyperparameters. Gonzalez and Kim [34] proposed an occlusion module and quantified disparity volume to improve geometric dependence. Their follow-up study [35] proposed neural positional encoding and distilled matting loss to improve the pixel-level depth estimation performance. However, the above methods have a problem in that a calibrated image pair must exist at training time.

Zhou et al. [36] proposed a network that simultaneously estimates depth and ego-motion from adjacent monocular images. They proposed viewpoint synthesis to reconstruct images with predicted camera poses and depths. They also used masks to improve the explainability of the model. Godard et al. [23] proposed the minimum reprojection loss using the minimum loss instead of the average in the calculation of photometric loss with adjacent images, which reduced the artificial structure of the image boundary and improved the sharpness of the occlusion boundary. They also proposed multiscale prediction to prevent the training target from being trapped in local gradient descent by binary sampling. Recent studies apply self-attention to global feature extraction [37], [38], [39], [40]. In addition, several studies improve performance by adding additional semantic information or motion information learning networks [41], [42], [43], [44]. In addition, multiframe-based evaluation is performed to further utilize geometric information [2], [37], [42], [45], [46].

Recent stereo-based depth estimation studies use cost volumes [47], [48], [49]. Im et al. [49] used warping transform to generate cost volume so that it can be used for unrectified stereo and multiimages. Recently, Watson et al. [2] proposed a cost-volume-based depth estimation that can be used in a sequence of images using pose estimation. They also proposed an adaptive cost volume, which made it possible to learn a minimum maximum depth from the data. In addition, they used a single image-based depth network as a teacher model, reducing errors for moving objects in cost volumes.

There is also a recent study using transformers for self-supervised depth estimation. Varma et al. [50] configured the network to learn camera parameters and compared the network according to the use of the CNN or transformer, respectively. Guizilini et al. [51] proposed a method to generate the cost volume from a cross-attention-based transformer network. However, it is necessary to use an additional network instead of the simple difference operation in previous studies.

Recent studies use a transformer for depth estimation and show good performance, but use a high amount of computation and parameters. In this study, we propose a hybrid network that is more efficient than existing CNN-based networks by hierarchically mixing CNNs and transformers.

III. METHOD

In this chapter, we describe the proposed hybrid transformer-based self-supervised learning depth estimation method. First, the viewpoint synthesis method of the self-supervised learning model and cost-volume-based depth estimation methodology are reviewed. This review describes the equations and geometric models used in the proposed model. Then, the hybrid encoder–decoder network proposed in this article will be described. The overall block diagram of the proposed structure is shown in Fig. 1.

A. View-Synthesis Model Based on Self-Supervised Learning

In this article, depth and pose networks are simultaneously trained for unsupervised learning-based depth estimation according to a recent study [2], [6], [23]. The network is trained through a view synthesis process that minimizes the photometric error between the target image \( I_t \) and the reconstructed target image \( \hat{I}_{s \rightarrow t} \) from the source image \( I_s \) to the target image viewpoint. The reconstructed image is sampled from the source image using the 2-D homogeneous coordinates obtained by projection using the predicted target depth and the predicted pose. At this time, the predicted depth \( D_t \) and pose \( P_{t \rightarrow s} \) are estimated by each network, and the camera parameter \( K \) is input. The viewpoint synthesis process for generating the reconstructed image is as follows:

\[
\hat{I}_{s \rightarrow t} = I_s \langle \text{proj}(D_t, P_{t \rightarrow s}, K) \rangle
\]

where proj is the camera projection operation and \( \langle \rangle \) is the binary sampling operation using STN [30].

The photometric error \( pe \) is combined with the \( L1 \) distance and SSIM [31], which is the degree of similarity. The image reconstruction loss is as follows:

\[
pe(I_t, \hat{I}_{s \rightarrow t}) = a \left( 1 - \text{SSIM}(I_t, \hat{I}_{s \rightarrow t}) \right) + (1 - a) \left\| I_t - \hat{I}_{s \rightarrow t} \right\|_1
\]

where \( a \) is the balancing weight and SSIM is a method of evaluating the quality of the images.

The source image consists of temporally adjacent frames of the target image. The reconstructed target image from the source image depends on the number of adjacent frames.
Camera movement or the presence of an occluded area are factors that increase image reconstruction loss. To reduce errors due to camera movement and occlusion area, a value with the smallest error is selected among several adjacent source images [23]. Equation (3) shows the formula for the minimum reconstruction loss error

\[ L_p = \min_S \left( I_t, \hat{I}_{s \rightarrow t} \right), \quad S \in [s_1, s_2, \ldots, s_n]. \]  

(3)

B. Cost-Volume-Based Depth Estimation

The recent cost-volume-based depth estimation of unsupervised learning proceeds in the following order. First, feature points are extracted from input images by a feature extractor. Next, the feature points of the source image are view-synthesized according to the assumed distance from the minimum to the maximum, and the cost volume is generated by calculating the similarity between the source and target feature points. The cost volume is aggregated in the encoder–decoder and converted to the disparity. Finally, view synthesis is performed with the predicted disparity. In this study, we perform cost volume-based depth estimation according to recent studies [2], [49].

The feature extractor that extracts the feature map from the input images uses the first two layers of ResNet. The feature extractor encodes the target image and the source image into the feature map viewpoint synthesis process. The feature map viewpoint synthesis formula is written as

\[ F_{s \rightarrow t}^d = F_S \langle \text{proj}(d, P_{t \rightarrow s}, K) \rangle, \quad d \in [d_{\text{min}}, d_{\text{max}}] \]  

(4)

where the symbols are the same as in (1). \( d_{\text{min}} \) and \( d_{\text{max}} \) are the minimum and maximum depths, respectively, and the step is adjusted according to the number of channels in the cost volume. To generate the cost volume, the above process is performed for each depth unit of the depth cost volume, and the \( L_1 \) distance of the feature map \( F_t \) for each unit and the reconstruction target feature map \( F_{s \rightarrow t}^d \) is input to each channel of the depth cost volume.

A cost-volume-based depth estimation that receives multiple frames generally works better than single frame estimation. However, the existence of an object moving in the same direction as the camera or a textureless region is a major cause of failure in cost-volume depth estimation. Because the cost-volume-based depth estimation uses the difference in feature points as a learning factor, the depth estimation fails when the depth difference cannot be known as above. To solve this problem, a recent study uses a single input image-based depth constraint network as a teacher model. The network used as the teacher model is the baseline of existing studies [6], [23]. The \( L_1 \) distance between the depth \( D_t \) predicted by the cost volume and the depth \( \hat{D}_t \) predicted by the depth constraint network is added to the loss function, preventing from excessively dependent on the disparity. The depth constraint loss is written as

\[ L_{\text{constraint}} = \| D_t - \hat{D}_t \|_1. \]  

(5)

Additionally, an edge-aware term such as (6) that constrains the gradient of depth according to the gradient of the image is added as in previous studies [4], [5], [6]

\[ L_{\text{smooth}} = |\delta_x D_t|e^{-|\delta_y D_t|} + |\delta_y D_t|e^{-|\delta_x D_t|}. \]  

(6)

The final loss consists of reconstruction loss, depth constraint loss, and depth smoothness loss and is as follows:

\[ L_{\text{final}} = \alpha L_p + \beta L_{\text{constraint}} + \gamma L_{\text{smooth}} \]  

(7)

where \( \alpha, \beta, \) and \( \gamma \) are loss function scale correction weights.

C. Hybrid Encoder and Self-Attention Decoder

In this article, we propose a novel encoder–decoder network that transforms cost volume into depth. The proposed network is efficient in terms of parameters and the computational amount and has high accuracy and low error metric.
The proposed encoder is constructed by hierarchically mixing networks with different characteristics to improve local and global feature representation performance. In the decoder, an attention operation layer is added to improve the global feature representation dependency.

1) Hybrid Encoder: The output of the feature extractor and the cost volume are concatenated. This concatenated output is input to the encoder–decoder network for depth estimation. The encoder reduces spatial resolution and extracts features at multiple scales.

Since the CNN-based model uses convolution operation, it has the characteristic of expressing features for the adjacent region, and the transformer has the characteristic of expressing the entire region with its self-attention mechanism. Therefore, in this study, we propose a hybrid encoder structure that hierarchically connects convolution and the transformer.

Convolution suitable for local feature expression is applied to the early layer of the encoder, and a transformer structure suitable for global feature expression is applied to the later layer by self-attention operation. The local feature block (LFB) is composed of the existing ResNet block. The global feature block (GFB) consists of a lightweight multiresolution module with halved keys, values, and an inverse residual feed-forward network [26]. Fig. 2 shows the detailed block diagram of the proposed hybrid encoder.

In detail, the LFB is composed of two residual blocks of existing ResNet, and the first residual block sets stride to 2 to reduce the resolution. When an input \( X \in \mathbb{R}^{H \times W \times C} \) is given, the LFB process is as follows:

\[
\text{LFB} (X) = \text{Residual} (\text{Residual} (X, s = 2)) \tag{8}
\]

where Residual is the existing ResNet block and \( s \) is the stride.

The GFB uses a CMT block that mixes convolution and the transformer [26]. The CMT block consists of a local perception unit (LPU), lightweight attention (LA), and IRFFN. First, the LPU for extracting local features is composed of \( 3 \times 3 \) depth-wise convolution and the sum of residuals. When an input feature \( X \) is given, the LPU is as follows:

\[
\text{LPU} (X) = \text{DWConv} (X) + X. \tag{9}
\]

Next, in the LA module, the 2-D input feature \( X \in \mathbb{R}^{H \times W \times C} \) is flattened to \( X \in \mathbb{R}^{N \times C} \) for patch operation. In this, \( N = H \times W \). To reduce the computational complexity of the attention operation, the key and the value reduce the spatial resolution with a \( k \times k \) depth-wise convolution set by stride \( k \), respectively. According to the existing self-attention operation, the query and the key are linearly transformed into the \( d_k \) dimension and the value into the \( d_v \) dimension. So, each is linearly transformed into a query \( Q \in \mathbb{R}^{N \times d_k} \), key \( K \in \mathbb{R}^{N \times (k^2) \times d_k} \), and value \( V \in \mathbb{R}^{N \times d_v} \). As in the recent study [16], a relative position bias \( B \) is added, and LA is as follows:

\[
\text{LA} (Q, K, V) = \text{Softmax} \left( \frac{QK^T}{\sqrt{d_k}} + B \right) V. \tag{10}
\]

IRFFNs are similar to the structure of MobilNetV2 [52], but the connection location of residuals has been changed. The modified IRFFNs are as follows:

\[
\text{IRFFN} (X) = \text{Conv} (\text{Conv} (X)) + X, \tag{11}
\]

Finally, the GFB is composed of each block and residual as follows:

\[
\text{GFB} (X) = L3 (L2 (L1 (X)))
\]

\[
L1 (X_{i-1}) = X’_{i} = \text{LPU} (X_{i-1})
\]

\[
L2 (X’_{i}) = X”_{i} = \text{LA} (\text{LN} (X’_{i})) + X’_{i}
\]

\[
L3 (X”_{i}) = X_{i} = \text{IRFFN} (\text{LN} (X”_{i})) + X”_{i} \tag{12}
\]

where \( X’_{i} \) and \( X”_{i} \) are the outputs of LPU and LA, respectively, and LN is layer normalization. Multiple GFBs are stacked in each stage.

Through the feature extractor and hybrid encoder, the spatial resolution of the entire feature map is reduced to \( F_{nx} \in \mathbb{R}^{(H/2^n) \times (W/2^n) \times C_n}, 1 \leq n \leq 5 \). Here, \( n \) is each layer number, composed of two layers in the feature extractor and three layers in the hybrid encoder.

2) Attention Decoder: In the decoder, the spatial resolution of the multiscale feature map is gradually restored, and the depth of the input resolution is estimated in units of pixels. First, self-attention operations are performed on low-resolution features in order to increase the wide feature expression and acceptance range [53]. The computed features are upscaled and merged with the higher-resolution feature maps. The detailed block diagram of the proposed self-attention decoder is shown in Fig. 3, and the input and output sizes and detailed structures of the entire network are shown in Table I.

The self-attention block consists of a channel attention operation and a spatial attention operation [53]. The self-attention operations consist of the 1-D channel self-attention operation \( M_c \in \mathbb{R}^{1 \times 1 \times C} \) and the 2-D spatial self-attention operation \( M_s \in \mathbb{R}^{H \times W \times 1} \) for the input feature \( X \in \mathbb{R}^{H \times W \times C} \). The total self-attention process is as follows:

\[
X’ = M_c (X) \otimes X
\]

\[
X” = M_s (X’) \otimes X’ \tag{13}
\]

where \( \otimes \) is element-wise multiplication, \( X’ \) is the result of the channel self-attention operation, and \( X” \) is the final result of the attention operation. After the self-attention operation, each
A. Experimental Setup

The experiment was performed in an NVIDIA RTX 3090 with a 24-GB memory hardware environment. The proposed hybrid encoder model was built based on the Manydepth model [2]. The depth and pose networks were trained with 40 epochs, a batch size of 8, and a $2 \times 10^{-4}$ learning rate. For weights, the ResNet layer is pretrained with ImageNet, and the transformer layer is trained from scratch. Unwritten parameters are used the same as the baseline study Manydepth.

B. Datasets

The KITTI dataset is data created for autonomous driving, captured by four cameras and lidar sensors, and synchronized with each other [54]. We follow the data distribution proposed by [55] with 39,810 for training, 4,424 for evaluation, and 697 for the test.

The Cityscapes dataset is a driving image dataset captured with a stereo camera [56]. Following [2], we use 69,731 images of the left sequence as training data. We use triples according to previous research. We evaluate the model with 1525 test images.

C. Evaluation Metrics

To compare the proposed depth estimation network with other modern networks, a commonly used evaluation metrics was used [2, 3, 4, 5, 6, 7, 13, 14, 29, 30, 31, 37, 38, 39, 40, 41, 42]. The five evaluation methods used in the experiment are AbsRel, SqRel, root mean square error (RMSE), RMSE (log), and accuracy index. In the evaluation methods, AbsRel, SqRel, and RMSE are error evaluation methods, the lower the better. Threshold accuracy is the accuracy evaluation metrics, where the higher the better.

D. Comparison Study

We compare the proposed method with the existing state-of-the-art methods. Table II shows the quantitative performance evaluation on the KITTI dataset of the existing depth estimation models and the proposed hybrid-based model. Test frames indicate the number of frames used in the test time, and the numbers in parentheses $-1$, $0$, and $1$ mean the previous frame, the current frame, and the next frame, respectively. The semantics column shows whether segmentation networks or semantic supervision are used. The T50 of the proposed model means that the backbone of a single-depth constraint network is changed from ResNet18 to ResNet50.

The proposed hybrid model shows higher performance in error metrics than the existing models. The proposed method shows higher performance than the single-frame-based estimation methods as well as the multiframe-based estimation methods. In the accuracy evaluation of $\delta < 1.25^3$, lower performance was recorded than the model [41] using semantic information and a heavier PackNet backbone network [6]. However, the difference is very small, and the proposed model showed high performance in most evaluation indicators without using semantic information. Furthermore, we found better performance when using ResNet50 as the backbone of a single-depth constrained network.

Table III shows comparative experiments on cityscapes datasets. Again, we show higher performance than previous studies in all evaluation methods.

![Structure of (a) attention block and (b) decoder block.](image)

### TABLE I

ARCHITECTURE OF THE PROPOSED NETWORK

| Module          | Input size : 640 x 192 x 3 | Architecture |
|-----------------|-----------------------------|-------------|
| Feature Extractor | 320 x 96 x 64               | 7 x 7 conv, stride 2 |
|                 | 160 x 48 x 64               | maxpooling, stride 2 |
| Encoder         | 80 x 24 x 128               | 3 x 3 conv, stride 2 |
|                 | 40 x 12 x 184               | Global Feature Block x 2 |
|                 | 20 x 6 x 368                | Global Feature Block x 2 |
| Decoder         | 40 x 12 x 256               | Attention Decoder Block |
|                 | 80 x 24 x 128               | 3 x 3 conv, stride 2 |
|                 | 160 x 48 x 64               | Global Feature Block x 2 |
|                 | 320 x 96 x 32               | 3 x 3 conv, stride 2 |
|                 | 640 x 192 x 16              | 3 x 3 conv, stride 2 |

Feature map is upsampled by a factor of 2 and concatenated with the feature map of the previous layer. The decoding process is as follows:

$$f_n^{\text{decode}} = \text{conv}_1 \left( \text{upsample}_{\times 2} \left( f_{n-1}^{\text{decode}} \right) + F_n \right)$$

$$\text{disp}_n = \sigma \left( \text{conv}_2 \left( f_n^{\text{decode}} \right) \right)$$

where $\text{disp}_n \in \mathbb{R}^{(H/2^n)\times(W/2^n)\times 1}$, $0 \leq n \leq 4$. Here, $n$ is the layer number, and $n = 0$ is the original resolution.
The resulting images for qualitative evaluation are shown in Fig. 4. Our method requires fewer parameters and computations than the existing models, but confirms that it shows similar depth estimation results. Specifically, we boxed flat areas such as a sign, truck, and bus. The proposed algorithm estimated the depth better in the boxed region than the existing algorithms. This is the result of estimating the depth of a flat area lacking local features based on regional features.

### E. Ablation Study

Table IV shows the results of the ablation study to evaluate the performance of each module of the proposed method. The basic model used the Manydepth model. Performance evaluation is performed according to the presence of the hybrid encoder and self-attention decoder proposed in this ablation study. The proposed hybrid encoder improves both error and accuracy evaluation indicators. In addition, the use of a hybrid encoder reduces parameters and computational amount. This means that compared to the ResNet layer, the lightweight transformer model maintains global feature expression performance with a low amount of computation. The self-attention decoder improves the SqRel and RMSE error metrics, but lowers the accuracy metrics. The use of all proposed modules improves both error and accuracy evaluation indicators. However, the performance of the accuracy evaluation metric is lower than when using only the hybrid encoder.
V. Conclusion

In this article, we proposed a hybrid model that combines the CNN and ViT for unsupervised learning-based monocular depth estimation. In the early layer of the encoder, the CNN is used to represent local features. The later layer of the encoder uses a transformer that extracts global features. We assumed that it is effective to have an expert model according to the depth of layer, and the performance improvement was confirmed through various experiments. In addition, high-computational problems of the transformer were solved by using an LA model, and the performance was improved by applying attention to the decoder.

As a result of comparative experiments on the KITTI dataset, the proposed method reduces AbsRel to 0.095, SqRel to 0.696, and RMSE to 4.317 and improves the accuracy index on $\delta < 1.25$ to 0.902. Also, the parameters are effectively reduced from 14.42 to 8.13 M and multiply-accumulate operations (MACs) are reduced from 13.77 to 12.23 G. The ablation study shows the performance improvement according to the use of the proposed module. Therefore, it is confirmed that the proposed hybrid transformer model is an efficient model that improves depth estimation performance.

Despite the improvement in the performance of the proposed study, there are still issues to be studied in the future. The proposed method has depth ambiguity through depth estimation based on image information. This causes a failure to accurately estimate the distance in meters. Some studies solve this problem by using IMU sensor data or external camera parameter information. In future research, it seems that this can be solved by network learning on the use of external data for absolute depth estimation.

References

[1] Y. Ming, X. Meng, C. Fan, and H. Yu, “Deep learning for monocular depth estimation: A review,” Neurocomputing, vol. 438, pp. 14–33, May 2021.

[2] J. Watson, O. M. Aodha, V. Prisacaru, G. Brostow, and M. Firman, “The temporal opportunistic; Self-supervised multi-frame monocular depth,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 1164–1174.

[3] X. Luo, J.-B. Huang, R. Szeliski, K. Matzen, and J. Kopf, “Consistent video depth estimation,” ACM Trans. Graph., vol. 39, no. 4, pp. 1–71, Aug. 2020.

[4] X.-Z. Cui, P. Feng, S.-Z. Wang, and J.-H. Zhang, “Monocular depth estimation with self-supervised learning for vineyard unmanned agricultural vehicle,” Sensors, vol. 22, no. 3, p. 721, Jan. 2022.

[5] J.S. Hwang, S.-J. Park, G.-M. Kim, and J.-H. Baek, “Unsupervised monocular depth estimation for colonoscopy system using feedback network,” Sensors, vol. 21, no. 8, p. 2691, Apr. 2021.

[6] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, “3D packing for self-supervised monocular depth estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 2485–2494.

[7] C. Shu, K. Yu, Z. Duan, and K. Yang, “Feature-metric loss for self-supervised learning of depth and egomotion,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Cham, Switzerland: Springer, Aug. 2020, pp. 572–588.

[8] H. Pham, Z. Dai, Q. Xie, and Q. V. Le, “Meta pseudo labels,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 11557–11568.

[9] X. Li et al., “Semantic flow for fast and accurate scene parsing,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Cham, Switzerland: Springer, Aug. 2020, pp. 775–785.

[10] M. Tan, R. Pang, and Q. V. Le, “EfficientDet: Scalable and efficient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10781–10790.

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[12] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 6105–6114.

[13] Z. Cheng, Y. Zhang, and C. Tang, “Swin-Depth: Using transformers and multi-scale fusion for monocular-based depth estimation,” IEEE Sensors J., vol. 21, no. 23, pp. 26912–26920, Dec. 2021.

[14] R. Rainfli, A. Bochkovskiy, and V. Koltun, “Vision transformers for dense prediction,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 12179–12188.

[15] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” 2020, arXiv:2010.11929.

[16] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted Windows,” 2021, arXiv:2103.14030.

[17] X. Dong et al., “C3DWin transformer: A general vision transformer backbone with cross-shaped Windows,” 2021, arXiv:2107.00652.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), vol. 25, Stateline, NV, USA, Dec. 2012, pp. 1097–1105.

[19] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556.

[20] A. Howard et al., “Searching for MobileNetV3,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1314–1324.

[21] X. Yao, Z. Zhu, C. Kang, S.-H. Wang, J. M. Gorriz, and Y.-D. Zhang, “AdaD-FNN for chest CT-based COVID-19 diagnosis,” IEEE Trans. Emerg. Topics Comput. Intell., early access, Jun. 1, 2022, doi: 10.1109/ETC.2021.3174869.

[22] X. Yao, X. Wang, S.-H. Wang, and Y.-D. Zhang, “A comprehensive survey on convolutional neural network in medical image analysis,” Multimedia Tools Appl., pp. 1–45, Aug. 2020.

[23] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, “Digging into self-supervised monocular depth estimation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 3828–3838.

[24] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 10347–10357.

[25] H. Wu et al., “CVT: Introducing convolutions to vision transformers,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 22–31.

[26] J. Grou et al., “CMT: Convolutional neural networks meet vision transformers,” 2021, arXiv:2107.06263.

[27] F. Khan, S. Salahuddin, and H. Javidnia, “Deep learning-based monocular depth estimation methods—A state-of-the-art review,” Sensors, vol. 20, no. 8, p. 2272, Apr. 2020.

[28] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, “Unsupervised CNN for single view depth estimation: Geometry to the rescue,” in Proc. Eur. Conf. Comput. Vis. Amsterdam, The Netherlands: Springer, 2016, pp. 740–756.
[29] C. Godard, O. M. Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 270–279.

[30] M. Jaderberg, K. Simonyan, and A. Zisserman, “Spatial transformer networks,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), vol. 28, 2015, pp. 1–9.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[32] F. Tosi, F. Aleotti, M. Poggio, and S. Mattoccia, “Learning monocular depth estimation infusing traditional stereo knowledge,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 9791–9801.

[33] J. Watson, M. Firman, G. Brostow, and D. Turkmukhametov, “Self-supervised monocular depth hints,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 2162–2171.

[34] J. L. Gonzalez and M. Kim, “Forget about the LiDAR: Self-supervised depth estimators with MED probability, volumes,” in Proc. Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 12626–12637.

[35] J. L. G. Bello 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 Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 6851–6860.

[36] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, “Unsupervised learning of depth and ego-motion from video,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1851–1858.

[37] J. Wang, G. Zhang, Z. Wu, X. Li, and L. Liu, “Self-supervised joint learning framework of depth estimation via implicit cues,” 2020, arXiv:2006.09876.

[38] A. Johnston and G. Carneiro, “Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 4756–4765.

[39] J. Yan, H. Zhao, P. Bu, and Y. Jin, “Channel-wise attention-based network for self-supervised monocular depth estimation,” in Proc. Int. Conf. 3D Vis., Dec. 2021, pp. 464–473.

[40] H. Zhou, D. Greenwood, and S. Taylor, “Self-supervised monocular depth estimation with internal feature fusion,” 2021, arXiv:2110.09482.

[41] V. Guizilini, R. Hou, J. Li, R. Ambrus, and A. Gaidon, “Semantically-guided representation learning for self-supervised monocular depth,” 2020, arXiv:2002.12319.

[42] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, “Unsupervised monocular depth and ego-motion learning with structure and semantics,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2019.

[43] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova, “Depth from video in the wild: Unsupervised monocular depth learning from unknown cameras,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8977–8986.

[44] H. Li, A. Gordon, H. Zhao, V. Casser, and A. Angelova, “Unsupervised monocular depth learning in dynamic scenes,” 2020, arXiv:2010.16404.

[45] P. Ruikamp, D. Gao, H. Chen, N. Navab, and B. Busam, “Attention meets geometry: Geometry guided spatial-temporal attention for consistent self-supervised monocular depth estimation,” in Proc. Int. Conf. 3D Vis., Dec. 2021, pp. 837–847.

[46] V. Patil, W. Van Gansbeke, D. Dai, and L. Van Gool, “Don’t forget the past: Recurrent depth estimation from monocular video,” IEEE Robot. Autom. Lett., vol. 5, no. 4, pp. 6813–6820, Oct. 2020.

[47] J. Zbontar and Y. LeCun, “Stereo matching by training a convolutional neural network to compare image patches,” J. Mach. Learn. Res., vol. 17, no. 1, pp. 2287–2318, Jun. 2016.

[48] J.-R. Chang and Y.-S. Chen, “Pyramid stereo matching network,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5410–5418.

[49] S. Im, H.-G. Jeon, S. Lim, and I. So Kweon, “DPSNet: End-to-end deep plane sweep stereo,” 2019, arXiv:1905.00538.

[50] A. Varma, H. Chawla, B. Zonoizi, and E. Arani, “Transformers in self-supervised monocular depth estimation with unknown camera intrinsics,” 2022, arXiv:2202.03131.

[51] V. Guizilini, R. Ambrus, D. Chen, S. Zakharov, and A. Gaidon, “Multi-frame self-supervised depth with transformers,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 160–170.

[52] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 4510–4520.

[53] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 3–19.

[54] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The KITTI dataset,” Int. J. Robot. Res., vol. 32, no. 11, pp. 1231–1237, 2013.

[55] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2014, pp. 2366–2374.

[56] M. Cords et al., “The cityscapes dataset for semantic urban scene understanding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 3213–3223.

Seung-Jun Hwang received the B.S. and M.S. degrees in electronics and information engineering from Korea Aerospace University (KAU), Goyang, Republic of Korea, in 2012 and 2014, respectively, where he is currently pursuing the Ph.D. degree with the School of Electronics and Information Engineering. His current research interests include image processing and computer vision.

Sung-Jun Park received the B.S. and M.S. degrees in electronics and information engineering from Korea Aerospace University (KAU), Goyang, Republic of Korea, in 2019 and 2021, respectively, where he is currently pursuing the Ph.D. degree with the School of Electronics and Information Engineering. His current research interests include image processing and computer vision.

Joong-Hwan Baek received the B.S. degree in telecommunication engineering from Korea Aerospace University, Goyang, Republic of Korea, in 1981, and the M.S. and Ph.D. degrees from Oklahoma State University, Stillwater, OK, USA, in 1987 and 1991, respectively. He was a Senior Researcher with the Electronics and Telecommunications Research Institute, Daejeon, South Korea, from 1991 to 1992. He has been a Professor with the School of Electronics and Information Engineering, Korea Aerospace University, since 1992. He is also the Director of the Video-Audio Space Convergence Technology Research Center, Korea Aerospace University. His research interests include image processing, computer vision, pattern recognition, and multimedia.

Byungkyu Kim received the Ph.D. degree in mechanical engineering from the University of Wisconsin-Madison, Madison, WI, USA, in 1997. From 1997 to 2000, he was a Technical Staff Member with the Center for X-ray Lithography, University of Wisconsin-Madison, where he developed a computer code for thermal modeling of a mask membrane and wafer during beam exposure. From 2000 to 2005, he worked for the Microsystems Center, Korea Institute of Science and Technology (KIST), Seoul, South Korea, as a Principal Research Scientist. He was an in-charge of developing a microcapsule-type robotic endoscope. He is currently a Professor with Korea Aerospace University, Goyang, Republic of Korea. His research interests include space mechanisms, robotics, micro/nano-manipulators, and biomedical application robots.