IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 7, 2020, accepted April 26, 2020, date of publication May 6, 2020, date of current version May 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2992815

Transfer2Depth: Dual Attention Network With Transfer Learning for Monocular Depth Estimation

CHIA-HUNG YEH[1,2], (Senior Member, IEEE), YAO-PAO HUANG[2], CHIH-YANG LIN[3], (Member, IEEE), AND CHUAN-YU CHANG[4], (Senior Member, IEEE)

[1]Department of Electrical Engineering, National Taiwan Normal University, Taipei 10610, Taiwan
[2]Department of Electrical Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan
[3]Department of Electrical Engineering, Yuan Ze University, Taoyuan 32003, Taiwan
[4]Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology, Douliu 64002, Taiwan

Corresponding author: Chih-Yang Lin (andrewlin@saturn.yzu.edu.tw)

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 108-2218-E-110-002-, Grant 108-2218-E-003-002-, Grant MOST 107-2218-E-003-003-, Grant MOST 107-2221-E-155-048-MY3, and Grant MOST 107-2218-E-110-004-.

ABSTRACT Monocular depth estimation poses a fundamental problem in many tasks. Although recent convolutional neural network-based methods can achieve high accuracy with very deep networks and complex architectures to exploit different cues and features, doing so not only increases the vulnerability of the model, but also increases the difficulty of convergence. Moreover, recent depth estimation methods for indoor environments are impractical for outdoor environments. In this work, we aim to develop a simple deep network structure to improve model effectiveness for depth estimation. We apply a dual attention module that can be inserted into any type of network to improve the power of representation, and additionally propose a training strategy which combines transfer learning and ordinal regression to improve training convergence. Even with a simple end-to-end encoder-decoder type of network architecture, we are able to achieve state-of-the-art performance on two of the biggest datasets for indoor and outdoor depth estimation: NYU Depth v2 and KITTI.

INDEX TERMS Computer vision, deep learning, monocular depth estimation, spatial-channel attention module, transfer learning.

I. INTRODUCTION

Estimating depth from a 2D image is a long-standing challenge in computer vision and scene understanding. A depth map can not only provide considerable help for 3D-related applications such as 3D object detection [1], scene segmentation [2], and simultaneous localization and mapping systems [3], but also help other research areas, such as image dehazing [4], image refocusing [5] and augmented reality [6].

Compared to depth estimation that uses multiple images [7]–[9], monocular depth estimation is more complicated. Depth estimation from a pure 2D image suffers from scale ambiguity problem because a 2D image can be generated from infinite combinations of object sizes and camera movement speeds. Several methods [10]–[12] have attempted to resolve this problem by extracting helpful features from a scene, such as textures, object sizes and occlusions. With the rise of convolutional neural networks in recent years, many works have introduced [13]–[20] CNNs into the task of depth estimation and have significantly improved performance. These works usually treat the monocular depth estimation problem as a regression problem and train the network with mean squared error loss or other point wise losses. Recently, in order to further improve the estimation accuracy, very deep and complex network architectures have been designed to exploit different types of features in a scene. This not only increases the vulnerability of the model, but also increases the difficulty of model training.

In this paper, we propose three strategies to improve depth estimation: 1) dual attention module, 2) transfer learning, and 3) incorporate ordinal regression for outdoor depth estimation. Snapshots from our results are presented in Fig. 1.

The associate editor coordinating the review of this manuscript and approving it for publication was Shiqi Wang.

FIGURE 1. Estimated depth using Transfer2Depth. Input image and our estimated depth map on KITTI dataset.

Motivated by human perceptions, people tend to selectively focus on a specific part of a scene, that is the concept of attention. We simulate this mechanism by enhancing the meaningful features in the feature patch with a specially designed spatial-channel dual attention module. The module in use is designed to have the same input and output shape. Thus, it can be inserted not only in our proposed network, but in any type of model as well to boost the quality of extracted features. This idea has been proved to be very effective. We are able to deliver a consistent improvement of around 5% in accuracy simply by applying our proposed module.

To further improve the estimation performance without increasing convolutions, which would result in greater memory costs and slower inference speed, we delve into research on methods to improve model convergence. Though most researchers train their models from scratch, Zamir *et al.* [21] indicated that many tasks are highly related and share similar features in their investigation. Motivated by [21], we introduce transfer learning into our training. We pretrain a high-performance model with the ImageNet [22] dataset, and then utilize the pre-trained weights to initialize our depth estimation training. By initializing our training with meaningful weights, we found our model converges significantly faster and requires less training data compared to previous state-of-the-art models.

Since the depth range of outdoor data is much wider than that of indoors, previous depth estimation networks cannot be directly applied to outdoor environments. In this paper, we incorporate the concept of ordinal regression from [17] into our training strategy. The continuous depth values are discretized into 80 intervals, and the regression training in depth estimation is cast as a multi-class classification, where an ordinal loss is introduced to train the network. The proposed network achieves state-of-the-art performance on two of the biggest benchmarks on indoor and outdoor depth estimation, i.e., NYU Depth v2 [23] and KITTI [24], reaching an improvement in performance of up to 10%.

The remainder of this paper is organized as follows. In Section 2, a brief review of related previous works is provided. We describe our proposed network, spatial-channel dual attention module and training strategy in Section 3. Besides quantitative and qualitative comparisons, several experiments analyzing the performance of different parts of our proposed network are provided in Section 4. Last but not least, we conclude the paper in Section 5

II. RELATED WORK

Depth estimation is a crucial problem within 3D computer vision, and how to resolve 3D structural information from 2D RGB images has been a very popular and important research topic. Traditional methods [25]–[27] utilize feature point movements to resolve geometric information from multiple images. Corresponding feature points between images are extracted and triangulation is utilized to estimate depth. Groundbreaking work from Saxena *et al.* [10] introduced machine learning to estimate the depth for 2D images with monocular cues. Since then, several approaches [11], [12], [28]–[31] following this concept with different representations have been introduced.

In recent years, due to the success convolutional neural networks have had in image understanding tasks, many works [1], [8], [9], [13]–[20] have proposed using CNN for depth estimation. Powerful deep architectures such as VGG, ResNet, and DenseNet have brought the accuracy of depth estimation to a new level. Recent works [32], [33] utilize multilayer deconvolutions to recover fine information. Skip connection design has been introduced in some encoder-decoder research [18], [19], [34] to preserve details from network inputs. Unsupervised or semi-supervised learning were recently introduced to the depth estimation problem [32], [34]–[37]. These methods usually utilize the estimated depth map to reconstruct a reference image from another image with a different view angle and build up disparity losses between the reference image and the reconstructed image to train the network.

Transfer learning has been proven to be very effective in different cases. Recently, Zamir *et al.* [21] investigated the relationship and modeled the transfer learning dependency of 26 tasks, 16 of which are 3D or geometric related topics. When Alhashim *et al.* [19] introduced this concept to the problem of depth estimation, transferring the model for object classification to depth estimation was highly effective.

Plug-in modules for convolutional neural networks are a newly emerging research topic. Recent research [17], [38], [39], [41] has typically designed special modules for specific tasks to improve a model. Some works aim to develop modules that can be inserted into any network without the need for any hyper parameter modifications. Wang *et al.* [38] developed a non-local module to resolve global feature relationships. Their module has since been ported into many different tasks for performance improvement. Attention type modules [39]–[42] focus on improving the quality of extracted features. Wang *et al.* [40] proposed an encoder-decoder type of attention module, which achieved good performance but is computationally expensive. Hu *et al.* [41] proposed using global average pooling to reduce the computational cost and exploit inter-channel relationships. Other research [42], [43] suggests that spatial attention is as important as channel attention for feature enhancement.

Previous state-of-the-art techniques for monocular depth estimation are described as follows. Eigen *et al.* [13] proposed a multi-scale network consisting of a coarse scale

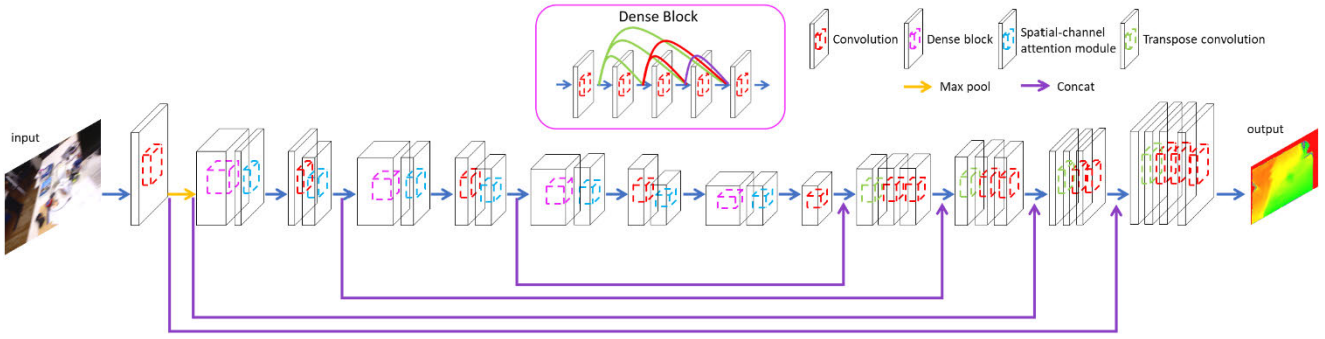

FIGURE 2. Overview of the network architecture. The network consists of an encoder-decoder pair with skip connections (purple), where our specially designed spatial channel dual attention module (blue) is inserted after each dense block (magenta) and bottle neck layer in the DenseNet architecture. Transpose convolutions (green) are utilized for spatial up-sampling in the decoder.

and a fine scale network. Even though the design goal of multi-scale network is to retain details while resolving global information, the pooling and striding at the beginning of the fine network causes it to lose a lot of information early on. In addition, unlike recently developed densely connected structures, which can preserve information while passing features, more details are lost after repeated convolution layers in the fine network. Fu *et al.* [17] regarded depth estimation as a multi-class classification problem. They used a dense feature extractor followed by a scene understanding module to extend the field of view to capture global information. However, the lack of connections between layers leads to a lack of detail in the output. Alhashim *et al.* [19] applied a simple encoder-decoder architecture and trained it using transfer learning technique. However, their architecture lacks attention to large scale features. Zhang *et al.* [20] designed a pattern affinitive network which concurrently produces depth, surface normal and semantic segmentation maps. The main hurdle in their approach is that the complex and massive architecture needs to be carefully engineered, which makes it fragile and increases the difficulty of convergence.

III. METHOD

In this section, we introduce the architecture and details of our proposed network.

A. NETWORK ARCHITECTURE

The overall network architecture is shown in Fig. 2. Previous depth estimation networks [13], [15] usually apply multiple layers of convolutional operations on different spatial sizes. However, as the architecture becomes deeper and deeper, the representation power of convolutional neural networks does not increase proportionally. Therefore, we incorporate a different aspect of network architecture called ''attention.'' Previous works have shown that attention not only tells the network where to focus, but also improves the representation of interest.

Encoder-decoder architecture [18], [19], [32]–[34] has been shown to be powerful in addressing the depth estimation problem. To preserve details, most works also introduce skip connections [18], [19], [33], [34] between the encoder and decoder. Our proposed network follows this trend as well. We incorporate high performance DenseNet [44] architecture and our proposed spatial-channel attention module as the encoder. The proposed spatial-channel attention module is inserted after each dense block and transition layer. For the decoder, we utilize transpose convolution for feature up sampling. Cross connections between the encoder and decoder are applied to preserve high level features for better output detail and quality.

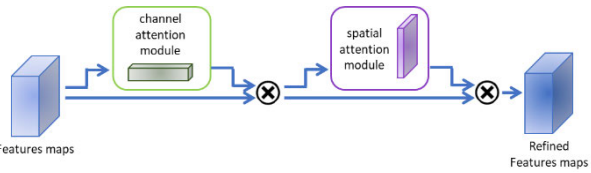

FIGURE 3. Spatial-channel dual attention module architecture. The module consists of two modules that work on different dimensions, where attention maps generated by each module is multiplied with input features to improve the representation power.

B. SPATIAL-CHANNEL DUAL ATTENTION MODULE

The spatial-channel dual attention module consists of two main components, a channel attention module and a spatial attention module. The overall architecture of the spatial-channel dual attention module is shown in Fig. 3.

The channel attention module is a combination of average pooling, max pooling and multi-layer perceptron, which is the same as in [43].

Because each channel of a feature map is seen as a distinct feature detector, channel attention focuses on finding out what is meaningful in the input feature pack. To preserve computational efficiency, average pooling and max pooling are applied to squeeze the spatial dimensions of the input feature map. Average pooling has been commonly adopted in previous works [41], [45]. However, in [43], with average pooling and max pooling both applied, the attention module is able to gather more important clues about distinctive object features. Thus, we simultaneously apply average pooling and max pooling.

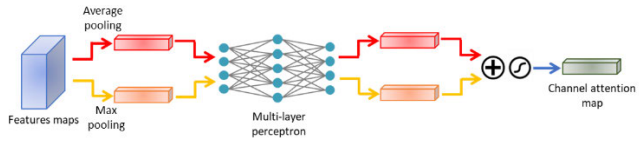

FIGURE 4. Channel attention module architecture. Same as in [43], the channel attention module is a combination of average pooling, max pooling and multi-layer perceptron.

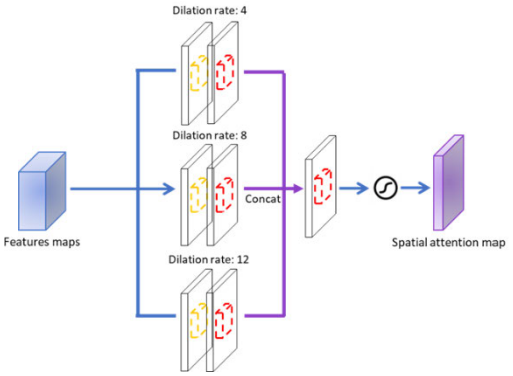

FIGURE 5. Spatial attention module architecture. Our proposed spatial attention module consists of an atrous spatial pooling pyramid followed by 1 × 1 convolutions.

As shown in Fig. 4, both the squeezed feature sequences from average pooling and max pooling are forwarded through a set of shared multi-layer perceptron. After the shared multi-layer perceptron is applied to each descriptor, the output features are merged using element-wise summation. A sigmoid activation is then applied to generate the channel attention sequence. The generated attention sequence multiplies with the input feature maps to emphasize the meaningful features.

In contrast to channel attention, which focuses on where the meaningful channels are, spatial attention focuses on the meaningful area in each given feature map. Unlike the spatial attention design in [43], our proposed spatial attention module consists of an atrous spatial pooling pyramid (ASPP) [46] followed by 1 × 1 convolutions. The ASPP extracts features from multiple receptive fields with dilated convolution operations. This avoids the loss of detail resulting from the spatial size reduction of features. After that, 1 × 1 convolutions learn the cross-channel interactions between the extracted features. We merge features from different receptive fields via concatenation, and a convolution with sigmoid activation is applied to further integrate and transform the features into a spatial attention map.

C. TRAINING STRATEGY

1) TRANSFER LEARNING

Alhashim *et al.* [19] show that with simple yet effective transfer learning technique, it is possible to significantly boost performance on the depth estimation problem. To further improve the performance of our proposed network, we therefore incorporate transfer learning to give our training a meaningful initialization. We pretrain a DenseNet169 dataset with ImageNet dataset. The pretrained weights are then transferred into our proposed network to initialize weight-setting, and the spatial-channel dual attention module and decoder are also inserted for depth estimation training.

2) ORDINAL REGRESSION

In most research, the depth estimation problem tends to be seen as a regression problem. However, the depth range of outdoor scenes is much wider than that of indoor scenes, so it is much more complex. Casting the depth estimation problem as a multiclass classification problem [17] can substantially simplify the problem, which leads to better estimation performance. Therefore, in this research we propose a split training strategy, where we switch our estimation method between regression and multiclass classification depending on the input data.

In order to perform ordinal regression for the outdoor environment, a ground truth depth map is first discretized into labels. We follow Hu *et al.* [17] in using a spacing-increasing discretization strategy, which avoids an over-strengthened loss for large depth values, letting our proposed network focus more on closer regions where 3D structural information is much richer than in the farther regions. Spacing-increasing strategy uniformly discretizes depth maps in log space, resulting in bigger intervals with larger depth values. The depth values are discretized into 80 classes, which are represented by 80 label maps. For each pixel, the labeling formula can be represented as:

$$t_i = e^x, \qquad x = \log(\alpha) + \frac{\log\left(\frac{\beta}{\alpha}\right)}{K}. \tag{1}$$

where $t_i \in \{t_1, t_2, \ldots, t_K\}$ are discretization thresholds. K is the number of intervals, which is set to 80 in this research. α and β are the shifted minimum and maximum depth values of the whole dataset, where we apply a shift ξ to both minimum and maximum depth values so that $\alpha = \text{minimum} + \xi = 1.0$.

Once the label of each pixel is decided, the label map is then filled in correspondingly. In label map representation, each label has its own map, which means there are 80 maps $L_i \in \{L_1, L_2, \ldots, L_{80}\}$ in this work. If a pixel $P_{(w,h)}$ is determined to be class 10, the corresponding pixel in label maps 1 through 10, $L_{1(w,h)}, L_{2(w,h)}, \ldots, L_{10(w,h)}$ is set to one while other areas and other maps remain zeros.

D. TRAINING AND INFERENCE

For indoor regression training, we use the loss function design from Alhashim *et al.*'s [19] work. The proposed loss function consists of three parts: point wise L1 loss, gradient loss and SSIM loss. The proposed loss function is outlined below:

$$L\left(d, \hat{d}\right) = \sigma \times L_{pointwise}\left(d, \hat{d}\right) + L_{grad}\left(d, \hat{d}\right) + \lambda \times L_{SSIM}\left(d, \hat{d}\right). \tag{2}$$

The point wise L1 loss is the average of the absolute error of each pixel to represent the overall disparity between the

estimated depth map and ground truth:

$$L_{pointwise}\left(d,\hat{d}\right)=\frac{1}{n}\sum_{p=1}^{n}\left|d_p-\hat{d}_p\right|. \tag{3}$$

The gradient loss is a L1 loss defined over the image gradient of the x and y axis:

$$L_{grad}\left(d,\hat{d}\right)=\frac{1}{n}\sum_{p=1}^{n}\left|g_x\left(d,\hat{d}\right)\right|+\left|g_y\left(d,\hat{d}\right)\right|, \tag{4}$$

where g_x represents the gradient of the depth map on the x axis and g_y represents the gradient of the depth map on the y axis. Since SSIM has an upper bound of 1 and a lower bound of 0, we define the SSIM loss as:

$$L_{SSIM}\left(d,\hat{d}\right)=1-SSIM\left(d,\hat{d}\right). \tag{5}$$

The weighting in the overall loss function is defined by $\sigma=0.1$ and $\lambda=0.5$, which is the same setting as in [19].

As for the outdoor dataset, we use Hu *et al.*'s [17] loss function design, which is an ordinal loss that takes the ordinal correlation between discrete labels into account. The ordinal loss function is formulated as the average of the pixelwise ordinal loss Ψ (w, h) over the entire image:

$$\begin{aligned}\mathrm{L}&=-\frac{1}{n}\sum_{w=0}^{W-1}\sum_{h=0}^{H-1}\Psi\left(w,h\right),\\ \Psi\left(w,h\right)&=\sum_{k=1}^{l_{(w,h)}}\log P^{k}_{(w,h)}+\sum_{k=l_{(w,h)}+1}^{K}\log\left(1-P^{k}_{(w,h)}\right).\end{aligned} \tag{6}$$

where k is the index of class labels, $\mathrm{P}^{\mathrm{k}}_{(\mathrm{w,h})}$ is the confidence of the predicted label of a pixel at position (w, h) and n is the total number of pixels in an image.

IV. EXPERIMENTAL RESULTS

In this section, we demonstrate the efficacy of our proposed method on two challenging datasets: NYU Depth v2 [23] and KITTI [24]. After introducing the implementation details, we compare our performance with state-of-the-art methods [17], [19], [20]. We follow previous work [13] on evaluation metrics and additionally perform ablation studies to further analyze the impact of different parts of our proposed method.

A. BENCHMARK PERFORMANCE

NYU Depth v2 The NYU Depth v2 is an indoor dataset with 464 indoor scenes of 640 × 480 resolution captured by a Microsoft Kinect depth camera. The dataset contains around 120K training samples and 654 testing samples pre-defined by previous work [13]. Just as in [19], a 50K image subset was selected as the training set. Since depth maps captured by Kinect usually contain a lot of invalid values, those invalid values are inpainted using the method in [47]. Depth maps in the NYU Depth v2 dataset have an upper bound of 10 meters.

KITTI The KITTI dataset is an outdoor dataset with about 1241 × 375 resolution captured by cameras and lidar sensors mounted on a moving vehicle. We train our network using 22.6K images as training images and 697 as testing images, following the settings in [13]. Ground truth resolution is reduced via max pooling for output measurements. We train our model with 640 × 480 resolution as the input, and 320 × 240 resolution as the output. Where we crop the input image to 640 × 375 and fill in zeros on the top to match the set input resolution.

B. IMPLEMENTATION DETAILS

We implement our proposed network with TensorFlow, and train on a Nvidia TITAN Xp GPU with 12 GB of memory. We pretrain a DenseNet 169 with ImageNet dataset for encoder weight initialization while the decoder weights are randomly initialized. We chose ADAM optimizer for our network with 20 epochs. The learning rate is set to 0.0001 with parameter $\beta_1=0.9$ and $\beta_2=0.999$, and the batch size is set to 2.

C. EVALUATION METRICS

We evaluate the proposed method's performance on six metrics with indoor data and seven metrics with outdoor data in line with previous work [13]. The error metrics for indoor data are defined as:

$$\mathrm{rel}=\frac{1}{n}\sum_{p=1}^{n}\frac{\left|d_p-\hat{d}_p\right|}{d_p}, \tag{7}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{p=1}^{n}\left(d_p-\hat{d}_p\right)^2}, \tag{8}$$

$$log_{10}=\frac{1}{n}\sum_{p=1}^{n}\left|\log_{10}d_p-\log_{10}\hat{d}_p\right|, \tag{9}$$

$$\delta_i=\%ofd_p\ satisfy\ \max\left(\frac{d_p}{\hat{d}_p},\frac{\hat{d}_p}{d_p}\right)<thr^i. \tag{10}$$

where $\mathrm{d_p}$ is a pixel in the ground truth depth map d and $\hat{d}_{\mathrm{p}}\hat{d}_{\mathrm{p}}$ is the corresponding depth value in the estimated depth map $\hat{d}$. n is the total number of pixels in each depth map. The seven metrics adopted for outdoor evaluation include rel, RMSE, and three threshold accuracies, as those adopted for indoor evaluation. The other two metrics are RMSElog and Squared Rel as follows:

$$\mathrm{logRMSE}=\sqrt{\frac{1}{n}\sum_{p=1}^{n}\left(\log d_p-\log\hat{d}_p\right)^2}, \tag{11}$$

$$\mathrm{sqr-rel}=\frac{1}{n}\sum_{p=1}^{n}\frac{\left\|d_p-\hat{d}_p\right\|^2}{d_p}. \tag{12}$$

D. PERFORMANCE

Table 1 shows the quantitative evaluation results on the NYU Depth v2 dataset. Our proposed method is able to

TABLE 1. Performance comparison on NYU Depth v2 dataset.

Method	higher is better			lower is better		
	δ_1	δ_2	δ_3	rel	log_{10}	RMSE
Make3D [12]	0.447	0.745	0.897	0.349	-	1.214
DepthTransfer [48]	-	-	-	0.350	0.131	1.200
Liu et al. [49]	-	-	-	0.335	0.127	1.060
Ladicky et al. [28]	0.542	0.829	0.941	-	-	-
Li et al. [50]	0.621	0.886	0.968	0.232	0.094	0.821
Wang et al. [14]	0.605	0.890	0.970	0.220	-	0.824
Roy et al. [51]	-	-	-	0.187	-	0.744
Liu et al. [16]	0.650	0.906	0.976	0.213	0.087	0.759
Eigen et al. [13]	0.769	0.950	0.988	0.158	-	0.641
Chakrabarti et al. [52]	0.806	0.958	0.987	0.149	-	0.620
Laina et al. [33]	0.811	0.953	0.988	0.127	0.055	0.573
Li et al. [53]	0.788	0.958	0.991	0.143	0.063	0.635
MS-CRF [54]	0.811	0.954	0.987	0.121	0.052	0.586
DenseDepth [19]	0.846	0.974	0.994	0.123	0.053	0.465
DORN [17]	0.828	0.965	0.992	**0.115**	0.051	0.509
Zhang et al. [20]	0.846	0.968	0.994	0.121	-	0.497
Transfer2Depth	**0.865**	**0.977**	**0.995**	0.117	**0.050**	**0.436**

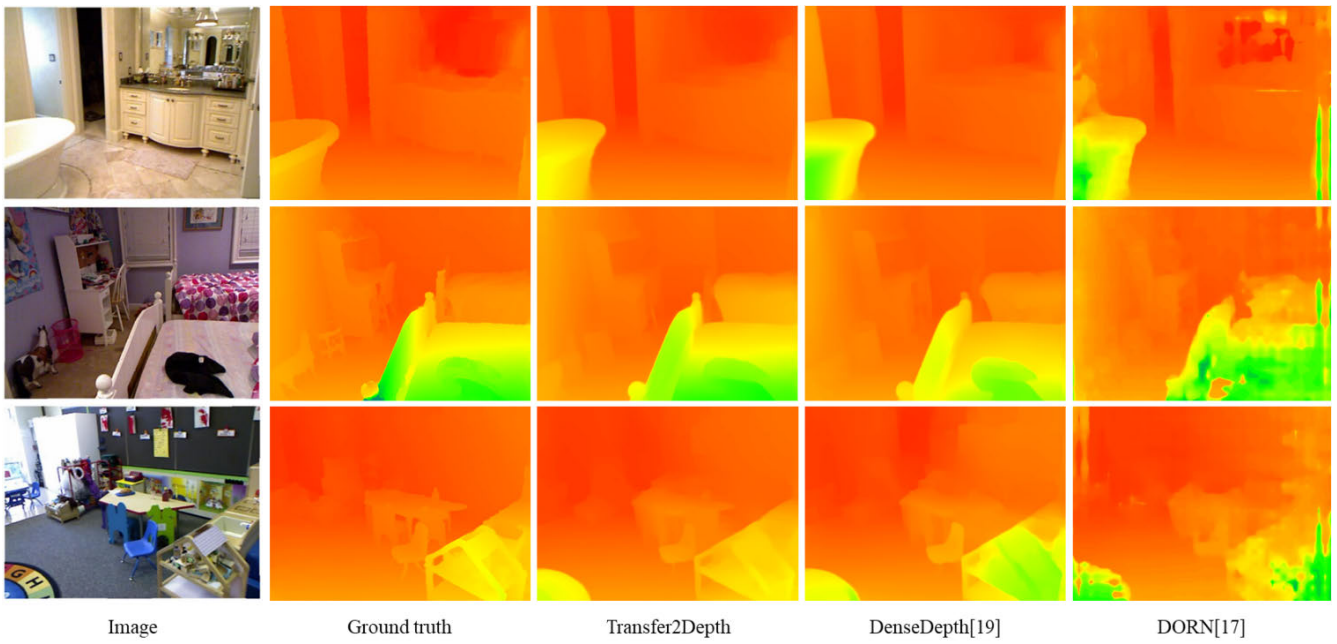

FIGURE 6. Depth prediction on NYU Depth v2 dataset. Input RGB image, ground truth depth map, our estimated depth map and depth map estimated by previous state-of-the-art [17], [19].

out-perform state-of-the-art methods and achieve up to a 6.2% improvement in performance while requiring only 50k training images and a 20 epoch training duration. The visualized qualitative comparison is shown in Fig. 6. As for outdoor data KITTI, although our quantitative comparison is slightly behind the previous best score [17] on squared relative error and RMSE, our proposed method achieves a 10% improvement on logRMSE as shown in Table 2. This indicates that our proposed method obtains better estimation at the closer range, which is much more important than the long range accuracy in most applications. This phenomenon may be derived from the pretraining process of our proposed network on ImageNet, which leads to better feature extraction on close range objects. On the other hand, our proposed method delivers significantly better qualitative results. As can be seen in Fig 7, our proposed method generates much sharper edges with smoother surfaces. These differences can be clearly observed on columnar objects such as road trees and traffic

TABLE 2. Performance comparison on KITTI dataset.

Method	higher is better			lower is better			
	δ_1	δ_2	δ_3	rel	sqr-rel	RMSE	logRMSE
Make3D [12]	0.601	0.820	0.926	0.280	3.012	8.734	0.361
Eigen et al. [13]	0.692	0.899	0.967	0.190	1.515	7.156	0.270
Liu et al. [16]	0.647	0.882	0.961	0.217	1.841	6.986	0.289
LRC [35]	0.861	0.949	0.976	0.114	0.898	4.935	0.206
Kuznietsov et al. [34]	0.862	0.960	0.986	0.113	0.741	4.621	0.189
DenseDepth [19]	0.886	0.965	0.986	0.093	0.589	4.170	0.171
DORN [17]	0.932	0.984	0.994	0.072	**0.307**	**2.727**	0.120
Transfer2Depth	**0.939**	**0.989**	**0.997**	**0.068**	0.333	3.178	**0.108**

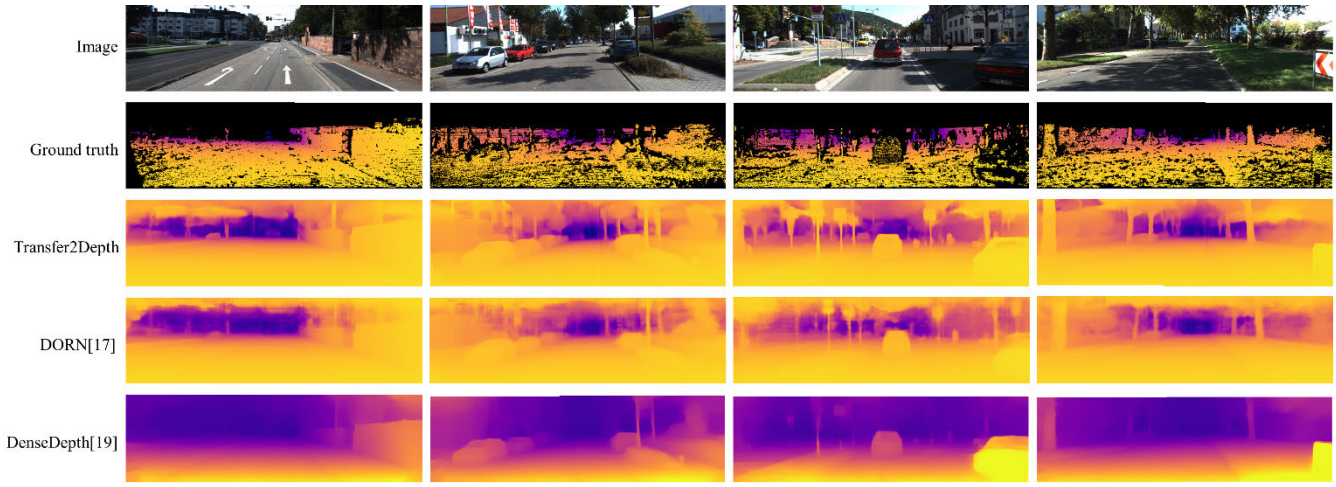

FIGURE 7. Depth prediction on KITTI. Input RGB image, ground truth depth map, our estimated depth map and depth map estimated by previous state-of-the-art [17]. Our method shows significant sharper edges and smoother surfaces.

TABLE 3. Spatial channel dual attention module.

Method	higher is better			lower is better		
	δ_1	δ_2	δ_3	rel	log_{10}	RMSE
w/o attention module	0.847	0.973	0.994	0.123	0.054	0.461
with Woo et al.'s module [43]	0.857	0.974	0.994	0.121	0.052	0.450
Transfer2Depth-atrous+pool	0.862	0.976	0.994	0.118	**0.050**	0.446
Transfer2Depth	**0.865**	**0.977**	**0.995**	**0.117**	**0.050**	**0.436**

sign poles. The results suggest that our proposed method provides state-of-the-art accuracy on both indoor and outdoor data.

E. ABLATION STUDIES

We performed several experiments to analyze the performance of different parts of our proposed network. All ablation study experiments were conducted on the NYU Depth v2 dataset.

1) SPATIAL ATTENTION MODULE DESIGN

Since the attention module plays a critical role in performance improvement, we run several experiments to prove its effectiveness and optimize the parameters. Table 3 shows the comparison of how different types of spatial attention module designs performed. With the application of our proposed attention module, we are able to gain around 5% improvement in accuracy compared to the same network architecture without the attention module. We also made the comparison among the performance of our proposed atrous convolution-only spatial design, the pooling only design of [43], and our spatial design when it is equipped with both atrous convolution and pooling. The evaluation results indicate that our proposed atrous convolution-based spatial attention module is able to extract features with better representation, which leads to better accuracy in depth estimation. The pooling operation in spatial attention is not practical because the max and average pooling lose significant amounts of spatial information.

TABLE 4. Transferred weights and network size.

Method	higher is better			lower is better		
	δ_1	δ_2	δ_3	rel	log_{10}	RMSE
DORN [17]	0.828	0.965	0.992	**0.115**	0.051	0.509
Zhang et al. [20]	0.846	0.968	0.994	0.121	-	0.497
Transfer2Depth-121	0.856	0.975	**0.995**	0.120	0.052	0.451
Transfer2Depth-169	**0.865**	**0.977**	**0.995**	0.117	**0.050**	**0.436**
Transfer2Depth-169 w/o t. l.	0.690	0.913	0.978	0.196	0.082	0.666

2) TRANSFERRED WEIGHTS AND NETWORK SIZE

In this experiment we test the influence of the transfer learning technique; also, we substitute the DenseNet-169 architecture for DenseNet-121 to test the performance of different encoder depth. Table 4 shows the comparison of performance with different weight initialization methods and encoder depths. The best performance at each depth is bolded and the second best is underlined. It can be seen that the impact of transfer learning technique is significant. As shown on the last row of Table 4, training without transfer learning leads to undesired performance due to the lack of training data and training epochs. By initializing our network with meaningful weights, we are able to gain a significant amount of improvement. In addition, even with a much smaller encoder, we are able to outperform the previous state-of-the-art techniques in most metrics. Though DenseNet offers a denser architecture with 201 layers, the previous work [19] argued that the performance improvement does not justify the trade-off with the much slower convergence and higher memory usage. Therefore, we conclude that utilizing DenseNet-169's architecture for our encoder achieves the best balance between performance and speed.

V. CONCLUSION

This paper proposes a convolutional neural network for monocular depth estimation from a single image. We leverage the effectiveness of high performing pre-trained models and a specially designed attention module. Unlike that most researches focus mainly on the network architecture design, our research aims to point out the importance of other aspects in model learning: training strategy and model effectiveness improvement. We propose a spatial-channel dual attention module which improves the representation power of the encoder, and a training strategy which combines transfer learning and ordinal regression to improve model convergence. Our proposed method can achieve a rate of 18 frames per second. Moreover, our results prove that our simple encoder-decoder module with attention function and ordinal regression is quite suitable for depth estimation in both indoor and outdoor environments using NYU Depth v2 and KITTI, two of the biggest datasets for indoor and outdoor images, respectively.

REFERENCES

[1] S. Song and J. Xiao, "Deep sliding shapes for amodal 3D object detection in RGB-D images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 808–816.
[2] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusionbased cnn architecture," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2016, pp. 213–228.
[3] R. Mur-Artal and J. D. Tardos, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.
[4] X. Ding, Y. Wang, J. Zhang, and X. Fu, "Underwater image dehaze using scene depth estimation with adaptive color correction," in *Proc. OCEANS Aberdeen*, Jun. 2017, pp. 1–5.
[5] F. Moreno-Noguer, P. N. Belhumeur, and S. K. Nayar, "Active refocusing of images and videos," *ACM Trans. Graph.*, vol. 26, no. 3, p. 67, Jul. 2007.
[6] W. Lee, N. Park, and W. Woo, "Depth-assisted real-time 3d object detection for augmented reality," in *Proc. Int. Conf. Artif. Reality Telexistence*, vol. 2, Nov. 2011, pp. 126–132.
[7] H. Ha, S. Im, J. Park, H.-G. Jeon, and I. S. Kweon, "High-quality depth from uncalibrated small motion clip," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5413–5421.
[8] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "DeMoN: Depth and motion network for learning monocular stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5622–5631.
[9] K. Wang and S. Shen, "MVDepthNet: Real-time multiview depth estimation neural network," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2018, pp. 248–257.
[10] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, 2006, pp. 1161–1168.
[11] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering surface layout from an image," *Int. J. Comput. Vis.*, vol. 75, no. 1, pp. 151–172, Jul. 2007.
[12] A. Saxena, M. Sun, and A. Y. Ng, "Make3D: Learning 3D scene structure from a single still image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, May 2009.
[13] 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.*, 2014, pp. 2366–2374.
[14] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. Yuille, "Towards unified depth and semantic prediction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2800–2809.
[15] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2650–2658.
[16] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2016.
[17] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2002–2011.
[18] Z. Hao, Y. Li, S. You, and F. Lu, "Detail preserving depth estimation from a single image using attention guided networks," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2018, pp. 304–313.
[19] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," 2018, *arXiv:1812.11941*. [Online]. Available: http://arxiv.org/abs/1812.11941
[20] Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe, and J. Yang, "Pattern-affinitive propagation across depth, surface normal and semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4106–4115.
[21] A. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3712–3722.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[23] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 746–760.

[24] 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, Sep. 2013.

[25] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 7–42, 2002.

[26] D. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.

[27] C. Wu, "Towards linear-time incremental structure from motion," in *Proc. Int. Conf. 3D Vis.*, Jun. 2013, pp. 127–134.

[28] L. Ladicky, J. Shi, and M. Pollefeys, "Pulling things out of perspective," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 89–96.

[29] S. Choi, D. Min, B. Ham, Y. Kim, C. Oh, and K. Sohn, "Depth analogy: Data-driven approach for single image depth estimation using gradient samples," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5953–5966, Dec. 2015.

[30] C. Hane, L. Ladicky, and M. Pollefeys, "Direction matters: Depth estimation with a surface normal classifier," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 381–389.

[31] R. Furukawa, R. Sagawa, and H. Kawasaki, "Depth estimation using structured light flow—Analysis of projected pattern flow on an Object's surface," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4640–4648.

[32] R. Garg, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 740–756.

[33] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 239–248.

[34] Y. Kuznietsov, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6647–6655.

[35] 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.

[36] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. M. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 340–349.

[37] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5667–5675.

[38] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[39] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[40] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.

[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.

[42] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5659–5667.

[43] 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.

[44] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[45] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[46] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[47] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Trans. Graph.*, vol. 23, no. 3, p. 689, Aug. 2004.

[48] K. Karsch, C. Liu, and S. B. Kang, "Depth transfer: Depth extraction from video using non-parametric sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2144–2158, Nov. 2014.

[49] M. Liu, M. Salzmann and X. He, "Discrete-continuous depth estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 716–723.

[50] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1119–1127.

[51] A. Roy and S. Todorovic, "Monocular depth estimation using neural regression forest," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5506–5514.

[52] A. Chakrabarti, J. Shao, and G. Shakhnarovich, "Depth from a single image by harmonizing overcomplete local network predictions," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 2658–2666.

[53] J. Li, R. Klein, and A. Yao, "A two-streamed network for estimating fine-scaled depth maps from single RGB images," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3372–3380.

[54] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5354–5362.

CHIA-HUNG YEH (Senior Member, IEEE) received the B.S. and Ph.D. degrees from the Department of Electrical Engineering, National Chung Cheng University, Chiayi, Taiwan, in 1997 and 2002, respectively. He was with the Department of Electrical Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan, as an Assistant Professor from 2007 to 2010, an Associate Professor from 2010 to 2013, and a Professor from 2013 to 2017. He is currently a Distinguished Professor with National Taiwan Normal University, Taipei, and the Vice Dean of College of Technology and Engineering. He has coauthored more than 250 technical international conferences and journal articles and held 47 patents in the USA, Taiwan, and China. His research interests include multimedia, video communication, three-dimensional reconstruction, video coding, image/video processing, and big data. He is a fellow of IET in 2017. He has been an Active TC Member of the IEEE Communication Society on Multimedia Communication, APSIPA, and IWAIT. He is also one of founding members of the ACM SIGMM Taiwan Chapter. He has been on the Best Paper Award Committee of JVCI and APSIPA. He was a recipient of the 2007 Young Researcher Award of NSYSU, the 2011 Distinguished Young Engineer Award from the Chinese Institute of Electrical Engineering, the 2013 Distinguished Young Researcher Award of NSYSU, the 2013 IEEE MMSP Top 10% Paper Award, the 2014 IEEE GCCE Outstanding Poster Award, the 2015 APSIPA Distinguished Lecturer, the 2016 NARLabs Technical Achievement Award, the Superior Achievement Award, the 2017 IEEE SPS Tainan Section Chair, the 2017 Distinguished Professor Award of NTNU, and the IEEE Outstanding Technical Achievement Award (the IEEE Tainan Section). He served as the Program Co-Chairs of the IEEE Big Data Multimedia 2016, IWAIT&IFMIA 2015, ICS 2014, ICICS 2013, APSIPA 2013, ICS2012, and the IEEE ISIC 2012, and the Co-Chair of the IEEE-TW 2016/2015, the IEEE ICME 2014, CVGIP2012, the IEEE PCM2012, VCIP 2012, APSIPA 2012, CVGIP2011, and VCIP 2010. He was an Associate Editor for the *Journal of Visual Communication and Image Representation*, the *EURASIP Journal on Advances in Signal Processing*, and the *APSIPA Transactions on Signal and Information Processing*.

YAO-PAO HUANG received the M.S. degree from the Department of Electrical Engineering, National Sun Yat-sen University, Taiwan, in 2019. His research interests include image/video processing, medical image processing, 2D/3D computer vision, pattern recognition, and deep learning.

CHIH-YANG LIN (Member, IEEE) was with the Advanced Technology Center, Industrial Technology Research Institute, Hsinchu, Taiwan, from 2007 to 2009. He was a Postdoctoral Fellow with the Institute of Information Science, Academia Sinica, Taipei, Taiwan, in 2009. He joined Asia University, Taichung, Taiwan, in 2010, where he was an Assistant Professor and then became an Associate Professor and a Professor, in 2013 and 2016, respectively. He was the Chair of the Department of Bioinformatics and Medical Engineering, Asia University, from August 2014 to January 2017. He is currently the Deputy Chief of the Global Affairs Office and a Professor with the Department of Electrical Engineering, Yuan-Ze University, Taoyuan, Taiwan. He has published over 100 articles in international conferences and journals with more than 1300 citations. He received the Best Paper Awards from the Pacific-Rim Conference on Multimedia (PCM), in 2008, the Best Paper Awards and the Excellent Paper Award from Computer Vision, Graphics and Image Processing Conference in 2009, 2013, and 2019, and the Best Paper Award from the 6th International Visual Informatics Conference 2019 (IVIC'19). He has served as a Session Chair, a Publication Chair, or a Workshop Organizer on many international conferences, including AHFE, ICCE, ACCV, IEEE Multimedia Big Data, ACM IH&MMSec, APSIPA, and CVGIP. He is also a regular Reviewer of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, IEEE ACCESS, and many other Elsevier journals. His research fields include computer vision, machine learning, deep learning, image processing, big data analysis, and the design of surveillance systems.

CHUAN-YU CHANG (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from National Cheng Kung University, Taiwan, in 2000. He was the Chair of Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology (YunTech), from 2009 to 2011. From 2011 to 2019, he served as the Dean of Research and Development, the Director of Incubation Center for Academia-Industry Collaboration and Intellectual Property (YunTech). He is currently the Chief Technology Officer (CTO) of Service Systems Technology Center, Industrial Technology Research Institute (ITRI), Taiwan. He is also a Distinguished Professor with the Department of Computer Science and Information Engineering, YunTech, Taiwan. His current research interests include computational intelligence and their applications to medical image processing, automated optical inspection, emotion recognition, and pattern recognition. In the above areas, he has more than 200 publications in journals and conference proceedings. He is an IET Fellow and a Life Member of IPPR and TAAI. He was the Chair of the IEEE Signal Processing Society Tainan Chapter and the Representative for Region 10 of the IEEE SPS Chapters Committee from 2015 to 2017. He served as the Program Co-Chair of TAAI 2007, CVGIP 2009, the 2010-2019 International Workshop on Intelligent Sensors and Smart Environments, and the third International Conference on Robot, Vision and Signal Processing (RVSP 2015). He served as the General Co-Chair of the 2012 International Conference on Information Security and Intelligent Control, the 2011-2013 Workshop on Digital Life Technologies, CVGIP2017, WIC2018, ICS2018, and WIC2019. He serves as an Associate Editor for two international journals including the *Multidimensional Systems and Signal Processing* and the *International Journal of Control Theory and Applications*. He is currently the President of Taiwan Association for Web Intelligence Consortium.