

# Attention-Aware Linear Depthwise Convolution for Single Image Super-Resolution

Seongmin Hwang, Gwanghyun Yu, Cheolkon Jung, *Member, IEEE*, and Jinyoung Kim

**Abstract**—Although deep convolutional neural networks (CNNs) have obtained outstanding performance in image super-resolution (SR), their computational cost increases geometrically as CNN models get deeper and wider. Meanwhile, the features of intermediate layers are treated equally across the channel, thus hindering the representational capability of CNNs. In this paper, we propose an attention-aware linear depthwise network to address the problems for single image SR, named ALDNet. Specifically, linear depthwise convolution allows CNN-based SR models to preserve useful information for reconstructing a super-resolved image while reducing computational burden. Furthermore, we design an attention-aware branch that enhances the representation ability of depthwise convolution layers by making full use of depthwise filter interdependency. Experiments on publicly available benchmark datasets show that ALDNet achieves superior performance to traditional depthwise separable convolutions in terms of quantitative measurements and visual quality.

**Index Terms**—Attention, convolutional neural network, determinant, deep learning, linear depthwise convolution, image super-resolution.

## I. INTRODUCTION

SINGLE image super-resolution (SISR) has gained much attention for the past decade. In general, the SISR task aims at restoring high-resolution (HR) image from its corresponding low-resolution input. However, SISR inherently has an ill-posed nature since a number of HR images can take the same low-resolution (LR) image by down-sampling, the solution space for SR problems is extremely large and there exist multiple HR solutions for LR input which makes SR operation an one-to-many mapping from LR image to HR image. To solve the ill-posed problem, a number of SR methods have been proposed for decades to produce plausible reconstructions by super-resolving high-frequency information from a given LR input, ranging from interpolation-based [1] to learning based methods [2], [3], [4], [5].

Among them, Dong et al. proposed SRCNN [2] and firstly demonstrated that CNN could be used to learn a mapping

This work was supported by the Ministry of Science and ICT (MSIT), Korea, under the High-Potential Individuals Global Training Program (2019-0-01609) supervised by the Institute for Information Communications Technology Planning Evaluation (IITP). Also, this work was supported by the National Natural Science Foundation of China (No. 61872280) and the International S&T Cooperation Program of China (No. 2014DFG12780).

S. Hwang, G. Yu, and J. Kim are with the School of Electronic and Computer Engineering, Chonnam National University, Yongbong-ro 77, Gwangju 61186, South Korea e-mail: 197209@jnu.ac.kr; sayney1004@naver.com; be-yondi@chonnam.ac.kr.

C. Jung is with the School of Electronic Engineering, Xidian University, Xi'an 710071, China e-mail: zhengzk@xidian.edu.cn.

(Corresponding authors: Cheolkon Jung and Jinyoung Kim).

LR image to HR image showing superior performance to traditional methods. Kim et al. [3] built a very deep network for SR (VDSR) using 20 layers by cascading small filters many times. Zhang et al. proposed the Residual Dense Network [5] (RDN) which fully used hierarchical features from all the convolutional layers by combining residual skip connections with dense connections.

Despite learning-based models, especially CNN-based models have achieved significant advances in image SR, and they are still facing two main problems: (1) CNN-based SR models suffer from computational complexity and memory consumption in practice as CNN models get deeper and wider to learn more discriminative high-level features. To overcome this problem, Howard et al. proposed depthwise separable convolution [6] that factorized a standard convolution into two convolutions: (1) depthwise convolution as a spatial filter, (2) pointwise convolution as a lightweight feature generator. Although depthwise separable convolution successfully achieves its superiority especially in terms of time and memory consumption, it is designed to solve high-level vision tasks such as image classification and object detection. Thus, depthwise separable convolution should be optimally adjusted prior to being applied to the low-level vision task. (2) Traditional CNN-based models usually adopt cascade network topologies. However, in this way the features in intermediate layers are all treated equally without considering inherent feature correlations. Hence, the features of each layer are sent to the sequential layer without any distinction, thereby hampering discriminative ability of CNN-based models.

To mitigate these drawbacks above, we propose an attention-aware linear depthwise network, named ALDNet, to preserve informative components as well as allow CNN-based SR models to have more powerful discriminative ability. The main idea of ALDNet is to remove non-linearity between depthwise convolution and pointwise convolution, thus making depthwise convolution linear mapping function. Linear depthwise convolution prevents non-linearity from destroying informative features, and thus helps reconstruct SR image. Furthermore, we build an attention-aware branch for depthwise convolution layer to adaptively recalibrate each channel-wise feature by modeling the interdependencies across all depthwise convolution filters. Such attention-aware branch makes full use of the advantage of depthwise convolution filter enabling ALDNet to strengthen more informative features. Hence, ALDNet is able to deal with the main problems that most of CNN-based models are facing. Experimental results demonstrate the superiority of ALDNet.

Compared with existing SR methods, the main contributions

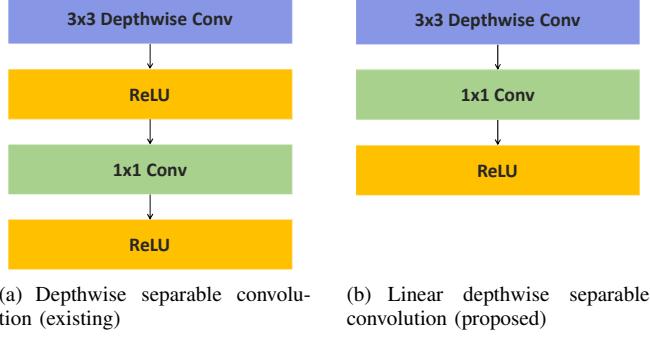


Fig. 1: Comparison between depthwise separable Convolution and linear depthwise separable convolution

of this paper are summarized as follows:

- We propose a novel ALDNet that learns correlations of depthwise convolution filters and remarkably enhances the discriminative capability of CNN models. To the best of our knowledge, this is the first work of using interdependencies among convolution filters to get attention.
- We present linear depthwise convolution to preserve informative features by removing non-linearity between depthwise convolution and pointwise convolution.
- We build a baseline network for SISR based on attention-aware linear depthwise blocks, named ALDSR. ALDSR, which is a very lightweight network, shows superior performance in SR to other state-of-the-art SR models.

## II. ATTENTION-AWARE LINEAR DEPTHWISE CONVOLUTION

### A. Linear Depthwise Convolution

Recently, MobileNet [6] proposed depthwise separable convolution and achieved outstanding performance in computer vision problems with low computational cost. However, depthwise separable convolution aimed to deal with high-level vision problems such as image classification. Thus, it is improper to apply depthwise separable convolution to low-level vision tasks such as image denoising and SR. Therefore, the MobileNet architecture should be modified to be used for image SR where the information should be handled more carefully. According to a former study [4] as shown in Fig. 1(a), we illustrated depthwise separable convolution by removing batch normalization [7] layers from standard depthwise separable convolution. Different from high-level vision tasks, since input and output domains of SR task are both images, the change in the distribution of network activations, so-called internal covariance shift [7], does not occur severely during training. Instead, batch normalization would rather get rid of network flexibility and lose scale information of an image while increasing GPU memory consumption during the training process. In this sense, batch normalization is no longer efficient for SR problem.

In this work, we propose linear depthwise convolution in which non-linearity (ReLU) is eliminated between depthwise convolutional layer and pointwise convolutional layer as

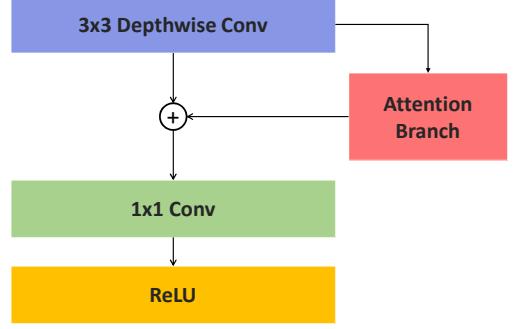


Fig. 2: Attention-aware linear depthwise convolution.

shown in Fig. 1(b). Depthwise separable convolution itself factorizes standard convolution by separating spatial filtering from the feature generation mechanism. Depthwise convolution, which acts as spatial filtering, performs convolution independently for every input channel with only considering spatial information.  $1 \times 1$  convolution, called pointwise convolution, computes linear combination considering channel information produced by depthwise convolution. That is, the feature map generated by depthwise convolution only considers spatial information.

We denote ReLU between depthwise convolution and pointwise convolution as "hasty ReLU". Consequentially, "hasty ReLU" is utilized by only considering spatial information without any consideration of channel features. Thus, "hasty ReLU" could cause destruction of information in SR reconstruction. Since the information preservation is a crucial factor for image SR, it is necessary to have special care on using non-linearity such as ReLU that causes information loss. Although "hasty ReLU" might work well on high-level vision task, it is obviously not good for solving low-level vision problems. As a result, we do not adopt "hasty ReLU" to achieve significant SR performance improvement.

### B. Attention-Aware Depthwise Convolution

Most previous CNN-based models generally treat the features in intermediate layers equally. To address this problem, SENet [8] was introduced to recalibrate the channelwise feature values in CNN-based models. In addition, recent works [9], [10] have shown that attention mechanism is helpful for improving SR performance to enhance more discriminative capability of CNNs. Typically, most of attention based CNN models adopt self-attention mechanism to get attention by capturing correlation between one pixel and other ones of the feature maps.

However, in CNN, of course, feature maps are important, but convolutional layers which directly generate feature values also play vital role. In this sense, we could consider utilizing parameters of convolutional filters for attention mechanism. Furthermore, since depthwise convolution filter has very few parameters compared with standard convolution, it is easier and more effective to extract representation values for getting attention. Added to this, depthwise convolution produces current feature maps by applying a single filter to each input channel unlike standard convolution, hence, the current feature

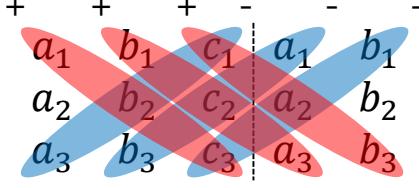


Fig. 3: Sarrus' rule to compute determinant.

maps are incomplete. However, if the network learns the importance of each depthwise filter and recalibrates the feature values by correlation of depthwise filters, the quality of the feature maps are improved, thus leading to a remarkable improvement of the network performance.

Driven by the observations, we propose attention-aware depthwise convolution which enables a network to learn importance of each depthwise filter of depthwise convolutional layer. By fully exploiting the interdependencies among depthwise filters, we improve representation power of CNN models significantly. We build a channel attention branch to model depthwise convolution interdependencies and recalibrate channelwise features as follows. The architecture of attention-aware linear depthwise convolution, which incorporates an attention-aware branch to linear depthwise convolution, is illustrated in Fig. 2.

**Determinant of depthwise filter:** Different from standard convolution, depthwise convolution has only single filter to produce feature maps, hence depthwise convolutional layer could be represented as a group of square matrices with the same number of rows and columns. Thus, we make full use of the distinctive feature of depthwise convolution as an informative descriptor. There are many methods to describe characteristics of square matrix. Among them, determinant captures important information about the matrix in a single number, and can be viewed as the volume scaling factor of the linear transformation by the square matrix. Since average and max pooling tend to oversimplify the filter information, we adopt determinant as the depthwise convolution descriptor.

Formally, given a group of depthwise filter  $k \times k \times C$ ,  $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_c]$  with  $C$  filters with kernel size of  $k$ . Each filter  $\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_c$  can be represented as  $k \times k$  square matrix as follows:

$$W_k^c = \begin{bmatrix} w_{1,1}^c & w_{1,2}^c & \cdots & w_{1,k}^c \\ w_{2,1}^c & w_{2,2}^c & \cdots & w_{2,k}^c \\ \vdots & \vdots & \ddots & \vdots \\ w_{k,1}^c & w_{k,2}^c & \cdots & w_{k,k}^c \end{bmatrix} \quad (1)$$

where  $w_{i,j}^c$  is a weight of convolution filter at position  $(i, j)$  of  $c$ -th channel.

In [11], Simonyan et al. reported that cascading  $3 \times 3$  filters has the same effect as the use of large filter sizes such as  $7 \times 7$  and  $11 \times 11$ , while reducing computational complexity. Thus, most of recent CNN-models adopt a combination of several convolutional layers with kernel size  $3 \times 3$ . The corresponding depthwise convolutional filter of a standard  $3 \times 3$  convolutional

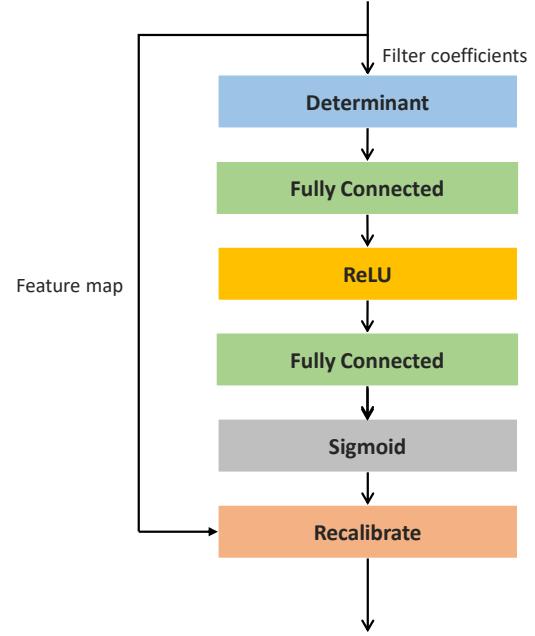


Fig. 4: Structure of attention branch.

filter is represented by

$$W_3^c = \begin{bmatrix} w_{1,1}^c & w_{1,2}^c & w_{1,3}^c \\ w_{2,1}^c & w_{2,2}^c & w_{2,3}^c \\ w_{3,1}^c & w_{3,2}^c & w_{3,3}^c \end{bmatrix} \quad (2)$$

where  $c$  stands for channel.

The rule of Sarrus is a well-known method to compute determinant of a  $3 \times 3$  matrix. As illustrated in Fig. 3, when the duplication of the first two columns of the matrix is written along with it, the determinant is calculated by subtracting the sum of the products of three diagonal southwest to north-east lines of elements from the sum of the products of three diagonal north-west to south-east lines of matrix elements. Consequentially, when the kernel size is 3, determinant captures correlation between diagonal components of the depthwise filter by sum of the products, thus determinant is able to capture shape of filter efficiently.

As the rule of Sarrus, the determinant of  $W_3^c$  is computed as follows:

$$\det W_3^c = w_{1,1}^c w_{2,2}^c w_{3,3}^c + w_{1,2}^c w_{2,3}^c w_{3,1}^c + w_{1,3}^c w_{2,1}^c w_{3,2}^c - w_{1,3}^c w_{2,2}^c w_{3,1}^c - w_{1,2}^c w_{2,1}^c w_{3,3}^c - w_{1,1}^c w_{2,3}^c w_{3,2}^c$$

With the help of determinant, we can shrink the depthwise filter to one dimensional vector  $\mathbf{z} \in \mathbb{R}^C$ .

**Attention branch:** In contradistinction to most generic CNN attention methods, we utilize interdependency among depthwise filters to get attention. To shrink the depthwise filter information, we take determinant on each depthwise convolution filter to produce one-dimensional vector  $\mathbf{z}$  that acts as a depthwise filter descriptor as mentioned earlier. To estimate attention across depthwise convolutional layer from the filter description vector  $\mathbf{z}$ , we opt to employ gating mechanism. As

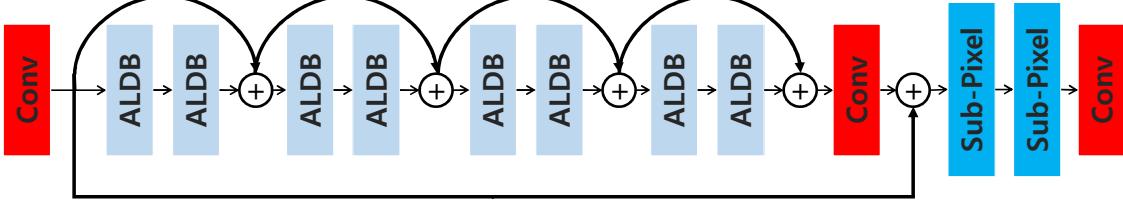


Fig. 5: Structure of the proposed ALDSR.

described in [8], an adequate gating function can be served by the sigmoid function.

$$\mathbf{s} = \sigma(\mathbf{W}_I \delta(\mathbf{W}_D \mathbf{z} + \mathbf{b}_D) + \mathbf{b}_I), \quad (3)$$

where  $\sigma$  and  $\delta$  are sigmoid function and rectified linear unit (ReLU), respectively;  $\mathbf{W}_I$  and  $\mathbf{W}_D$  are the weight set of convolutional layers;  $\mathbf{b}_I$  and  $\mathbf{b}_D$  are the corresponding biases. To avoid a parameter overhead, the ReLU activation size is set to  $\mathbf{z} \in \mathbb{R}^{C/r \times 1 \times 1}$ , where  $r$  is the reduction ratio. We set the reduction ratio  $r$  to 16. Then, we obtain the channel-attention map  $\mathbf{s}$  to rescale the feature map  $\mathbf{f}$  as follows:

$$\mathbf{f}'_c = s_c \cdot \mathbf{f}_c \quad (4)$$

According to [12], stacking attention modules naively leads to hinder the performance by dot production with mask range from zero to one repeatedly. Thus, we counteract this effect by adopting a residual learning strategy [13] to make network stable.

$$\begin{aligned} \mathbf{D}_c &= \mathbf{f}_c + \mathbf{f}'_c & (5) \\ &= (1 + s_c) \cdot \mathbf{f}_c, & (6) \end{aligned}$$

where  $\mathbf{D}$  represents the output of attention-aware depthwise convolution.

### III. ATTENTION-AWARE LINEAR DEPTHWISE NETWORK FOR IMAGE SR

#### A. Network Architecture

In this work, we propose a baseline network for image SR based on the attention-aware linear depthwise block (ALDB), named ALDSR. The overall architecture of ALDSR is illustrated in Fig. 5, which is constructed with the proposed attention-aware linear depthwise block (ALDB). Given  $I_{LR}$ , ALDSR generates  $I_{SR}$  where  $I_{LR}$  and  $I_{SR}$  stand for low-resolution image and its super-resolved counterpart, respectively. As shown in the figure, two types of residual learning are used to construct the network: (1) global residual learning [3] to provide skip-connection in global scale, and (2) local residual learning to provide skip-connection in every two ALDBs.

According to the former studies [5], [10], [9], we design ALDSR that mainly consists of four parts: shallow feature extraction, deep feature extraction via ALDB, upsample net, and reconstruction part. In ALDSR, only one convolutional layer is used to extract shallow feature  $\mathbf{F}_0$  from  $I_{LR}$  as follows:

$$\mathbf{F}_0 = H_{SF}(\mathbf{I}_{LR}), \quad (7)$$

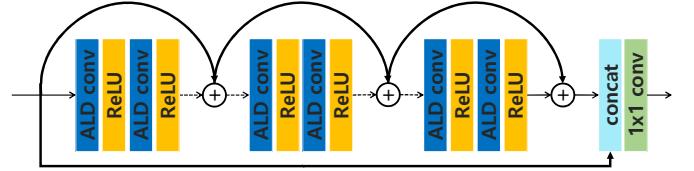


Fig. 6: Structure of the attention-aware linear depthwise block (ALDB).

where  $H_{SF}$  is convolution operation. Then,  $\mathbf{F}_0$  is used for global residual learning and deep feature extraction with ALDBs as follows:

$$\mathbf{F}_{DF} = H_{ALDBs}(\mathbf{F}_0), \quad (8)$$

where  $H_{ALDBs}$  is a deep feature extraction structure which consists of several ALDB. Next,  $\mathbf{F}_{DF}$  is upscaled by upsample net as follows:

$$\mathbf{F}_{UP} = H_{UP}(\mathbf{F}_{DF}), \quad (9)$$

where  $\mathbf{F}_{UP}$  stands for upsample net. As explored in [4], [5], ESPCN [14] is used to increase the spatial dimensions of the feature maps. Thus, the upscaled features are passed through one convolutional layer to be mapped into SR image as follows:

$$\mathbf{I}_{SR} = H_R(\mathbf{F}_{UP}) \quad (10)$$

$$= H_{ALDSR}(\mathbf{I}_{LR}), \quad (11)$$

where  $H_R$ ,  $H_{ALDSR}$  denote the reconstruction layer and function of proposed ALDSR, respectively. ALDSR is optimized with a  $L_1$  loss function which has been demonstrated to be more powerful for performance and convergence [4].

#### B. Attention-Aware Linear Depthwise Block

As shown in Fig. 5, ALDSR heavily relies on ALDBs. After applying a  $3 \times 3$  single convolution layer to the input LR images for learning shallow features, a set of ALDBs is employed to learn more discriminative features. We describe the details of the proposed ALDB illustrated in Fig. 6. Our ALDB is mainly constructed by cascading several attention-aware linear depthwise separable convolution (ALD convolution) with residual learning. Each ALD convolution consists of attention-aware linear depthwise convolutions followed by  $1 \times 1$  pointwise convolution and ReLU. The pointwise convolution layer is omitted for simplicity in Fig. 6.

TABLE I: ALDSR performance on public benchmark test datasets and DIV2K validation dataset in terms of PSNR(dB) and SSIM).

Dataset	Bicubic	VDSR [3]	LapSRN [15]	MemNet [16]	IDN [17]	ALDSR
Set5	28.43 / 0.8109	31.35 / 0.8838	31.54 / 0.8866	31.74 / 0.8893	<b>31.82 / 0.8903</b>	31.78 / 0.8895
Set14	26.00 / 0.7023	28.02 / 0.7678	28.09 / 0.7694	28.26 / 0.7723	28.25 / 0.7730	<b>28.37 / 0.7760</b>
B100	25.96 / 0.6678	27.29 / 0.7252	27.32 / 0.7264	27.40 / 0.7281	<b>27.41 / 0.7297</b>	<b>27.41 / 0.7304</b>
Urban100	23.14 / 0.6574	25.18 / 0.7525	25.21 / 0.7553	25.50 / 0.7630	25.41 / 0.7632	<b>25.55 / 0.7685</b>
DIV2K validation	28.11 / 0.775	29.82 / 0.824	29.88 / 0.825	-	-	<b>30.16 / 0.8313</b>
Parameters	-	665k	812k	677k	796k	731k

TABLE II: ALD-RDN performance on public benchmark test datasets and DIV2K validation dataset in terms of PSNR(dB) and SSIM.

Dataset	RDN [5]	RDN(re-implemented)	DW-RDN	ALD-RDN
Set5	32.47 / 0.8990	32.40 / 0.8977	32.10 / 0.8937	<b>32.18 / 0.8954</b>
Set14	28.81 / 0.7871	28.73 / 0.7861	28.59 / 0.7812	<b>28.62 / 0.7837</b>
B100	27.72 / 0.7419	27.68 / 0.7397	27.57 / 0.7354	<b>27.61 / 0.7378</b>
Urban100	26.61 / 0.8028	26.49 / 0.7988	26.06 / 0.7843	<b>26.23 / 0.7914</b>
DIV2K validation	-	30.65 / 0.8430	30.45 / 0.8375	<b>30.54 / 0.8406</b>

Similar to the ALDSR architecture, local residual learning is also employed in ALDB by adding short-path skip-connection in every two ALD convolutions. Furthermore, we adopt contiguous memory mechanism [5] which reads state from the preceding building block to make full use of hierarchical features that is able to get more information for SR reconstruction. However, it puts high computational burden on the network, thus it is required to simplify it. The modified contiguous memory mechanism is realized by passing the state of the preceding ALDB only to  $1 \times 1$  bottleneck layer of current ALDB. Then, it is concatenated with the final residual output of current ALDB. Bottleneck layer which uses  $1 \times 1$  convolution is used to adaptively fuse the concatenated features and reduce the channel size to half.

#### IV. EXPERIMENTS

##### A. Experimental Setup

The DIV2K [18] dataset consists of 800 high-resolution images for training, 100 images for validation, and 100 images for test. Since the ground truth in the test images is not released, we compare performance on the 100 validation images. We also use four standard benchmark datasets: Set5 [19], Set14 [20], BSD100 [21], and Urban100 [22]. For experiments, we generate LR images by  $4 \times$  bicubic-downsampling from HR images. We evaluate SR results in Y channel of the transformed YCbCr space in terms of peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). For training, 16 LR color patches of size  $48 \times 48$  from LR images with the corresponding HR patches provided as inputs. We augment the patches by randomly flipping horizontally or vertically and rotating  $90^\circ$ . We adopt Adam Optimizer [23] for training with the initial learning rate of  $10^{-4}$  and halved once at epoch 200. We train all models by 300 epochs. All experiments are implemented on PyTorch framework [24].

##### B. Performance Evaluation and Analysis

**ALD:** We first compare the proposed ALDSR with other state-of-the-art SR methods including VDSR [3], LapSRN [15], MemNet [16], IDN [17]. Table I shows the average

TABLE III: Comparison among building blocks in terms of the number of parameters.

Building block	Number of parameters
RDB	1,363,968
DW-RDB	205,056
LDW-RDB	205,056
ALD-RDB	257,280
ALDB	41,472

PSNR and SSIM values on four benchmark datasets and DIV2K validation dataset. Our ALDSR exhibits a significant improvement compared to the other methods on most datasets. The gaps is obviously noticeable on Set14 with outperforming MemNet [16] by 0.11dB.

To further investigate the effectiveness of the proposed ALDNet, we apply attention-aware linear depthwise convolution to a state-of-the-art architecture, Residual Dense Network [5], which is constructed to cascade several building blocks so-called RDB (Residual Dense Block). We construct ALDNet equivalent of RDN by simply replacing standard convolutional layers in RDB to ALD convolutional layers (Fig. 2). Furthermore, we also construct depthwise(DW)-RDN where standard convolutional layers are replaced with depthwise separable convolutional layers (Fig. 1(a)) in RDB.

The benchmark results of ALD-RDN are reported in Table II, which shows the overall average PSNR and SSIM. As shown in Table II, compared with DW-RDN, our ALD-RDN produces better results for all benchmark datasets. Especially, on the Urban100 dataset which contains images with complex patterns, ALD-RDN shows significant performance improvement, which indicates ALD-RDN is very effective for image SR with complex patterns. Since RDN requires quite higher computational complexity than DW-RDN and ALD-RDN, the original RDN achieves the best performance. Also, since the original RDN is trained over 300 epochs, it obtains better results than reimplemented version. The number of parameters of building blocks among them is reported in Table III.

**Effect of linearity:** To verify whether the linear depthwise convolution performs a crucial role in performance, we construct LDW-RDN. Specifically, we remove "ReLU" between

TABLE IV: Performance comparison by using different descriptors.

Dataset	ALDSR			ALD-RDN		
	Average	Max	Determinant	Average	Max	Determinant
Set5	31.77 / 0.8892	31.76 / 0.8892	<b>31.78 / 0.8895</b>	32.20 / 0.8953	<b>32.23 / 0.8957</b>	32.18 / 0.8954
Set14	<b>28.37</b> / 0.7758	28.36 / 0.7757	<b>28.37 / 0.7760</b>	<b>28.66 / 0.7837</b>	28.63 / 0.7827	28.62 / <b>0.7837</b>
B100	<b>27.41</b> / 0.7302	<b>27.41</b> / 0.7302	<b>27.41 / 0.7304</b>	27.60 / 0.7375	<b>27.61 / 0.7372</b>	<b>27.61 / 0.7378</b>
Urban100	<b>25.56</b> / 0.7680	<b>25.56</b> / 0.7680	25.55 / <b>0.7685</b>	26.20 / 0.7905	<b>26.23 / 0.7908</b>	<b>26.23 / 0.7914</b>
DIV2K validation	<b>30.17 / 0.8313</b>	30.16 / 0.8311	30.16 / <b>0.8313</b>	30.52 / 0.8402	<b>30.54 / 0.8400</b>	<b>30.54 / 0.8406</b>

TABLE V: Performance comparison between depthwise convolution and linear depthwise convolution.

Dataset	DW-RDN	LDW-RDN
Set5	32.10 / 0.8937	<b>32.17 / 0.8950</b>
Set14	28.59 / 0.7812	<b>28.60 / 0.7826</b>
B100	27.57 / 0.7354	<b>27.59 / 0.7367</b>
Urban100	26.06 / 0.7843	<b>26.14 / 0.7883</b>
DIV2K validation	30.45 / 0.8375	<b>30.50 / 0.8393</b>

depthwise convolution and pointwise convolution in DW-RDN, and the corresponding results are reported in Table V. We observe that linear depthwise convolution improves performance remarkably without any additional parameters. Especially, LDW-RDN yields a performance improvement from 26.06dB to 26.14dB on the Urban100 dataset which contains complex images and is difficult to reconstructed. Thus, linear depthwise convolution prevents destruction of informative features to reconstruct SR image from non-linearity such as "ReLU", and thus it performs better on the images with complex structures.

**Effect of determinant descriptor:** We also examine the effect of the determinant descriptor. To that end, we select average and max pooling to describe depthwise filter characteristics and conduct experiments with ALDSR and ALD-RDN. The results are reported in Table IV. Although all descriptors perform well on ALDSR, the average descriptor achieves slightly better PSNR performance than the others in most datasets. On the other hand, on ALD-RDN, max pooling generally produces better PSNR results. Determinant descriptor, which is used for ALDNet, exhibits the best SSIM results in all benchmark datasets except for Set5 on ALD-RDN, as well as it obtains comparable PSNR results to the others.

**Discussion:** Unlike standard convolution, depthwise convolution needs very few parameters. Since the maximum value of them represents a value with the most dominant influence on the convolution transform, it is effective to describe the convolution which has few parameters by max pooling. Thus, max pooling shows better PSNR results. However, max pooling describes only one value in the depthwise filter, and thus it is difficult to capture any information on describing the entire parameters of filter by max pooling. For this reason, max pooling shows unfavorable SSIM results. On the other hand, since determinant considers all the parameters carefully in the depthwise filter enabling network to learn importance of depthwise filter based on a lot of information of filter including shape, it can be found that not only the PSNR is high but also the SSIM is significantly higher.

**Visual Quality:** We provide visual comparison among ALDSR and state-of-the-art SR methods in Fig. 7. For challenging details in images "img004" from Urban100, most SR methods suffer from undesirable artifacts and heavy blurry results, however the proposed ALDSR generates good SR image. In addition, severe distortions are found in some reconstruction results by other SR methods, whereas ALDSR can reconstruct the repetitive patterns well without severe distortions. We also provide visual comparison among RDN, DW-RDN, LDW-RDN and ALD-RDN in Fig. 8. It can be observed that the proposed ALD-RDN exhibits comparable visual quality to RDN and even achieves better performance in lattice. In "img078" from Urban100, bicubic interpolation loses details and texture resulting in a very blurry image. DW-RDN is able to recover edges and coarse details, but it still fails to get fine details with wrong lattice. LDW-RDN produces better reconstruction results than DW-RDN. Compared with the ground-truth, our ALD-RDN reconstructs more convincing SR images with accurate lattice and fine details even obtaining better visual quality than RDN. It can be observed that determinant descriptor shows superior performance to average and max descriptors. These observations ensure the effectiveness of determinant-based ALDNet with powerful representation ability and reconstruction performance by informative feature preservation.

## V. CONCLUSION

In this paper, we have proposed an attention-aware linear depthwise network (ALDNet) for image SR that preserves informative features and enhances representation ability while reducing computational burden. Linear depthwise convolution allows ALDNet to prevent destruction of useful information that provides clues for SR reconstruction. Moreover, attention branch on linear depthwise convolution layer learns importance of each depthwise convolution filter, and thus allows ALDNet to strengthen informative features. To facilitate depthwise convolution, we use determinant as a linear depthwise convolution descriptor because the depthwise filter can be represented by square matrix. Experimental results demonstrate that ALDNet achieves state-of-the-art SR performance with low computational cost.

## REFERENCES

- [1] L. Zhang and X. Wu, "An edge-guided image interpolation algorithm via directional filtering and data fusion," *IEEE transactions on Image Processing*, vol. 15, no. 8, pp. 2226–2238, 2006.
- [2] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*. Springer, 2014, pp. 184–199.

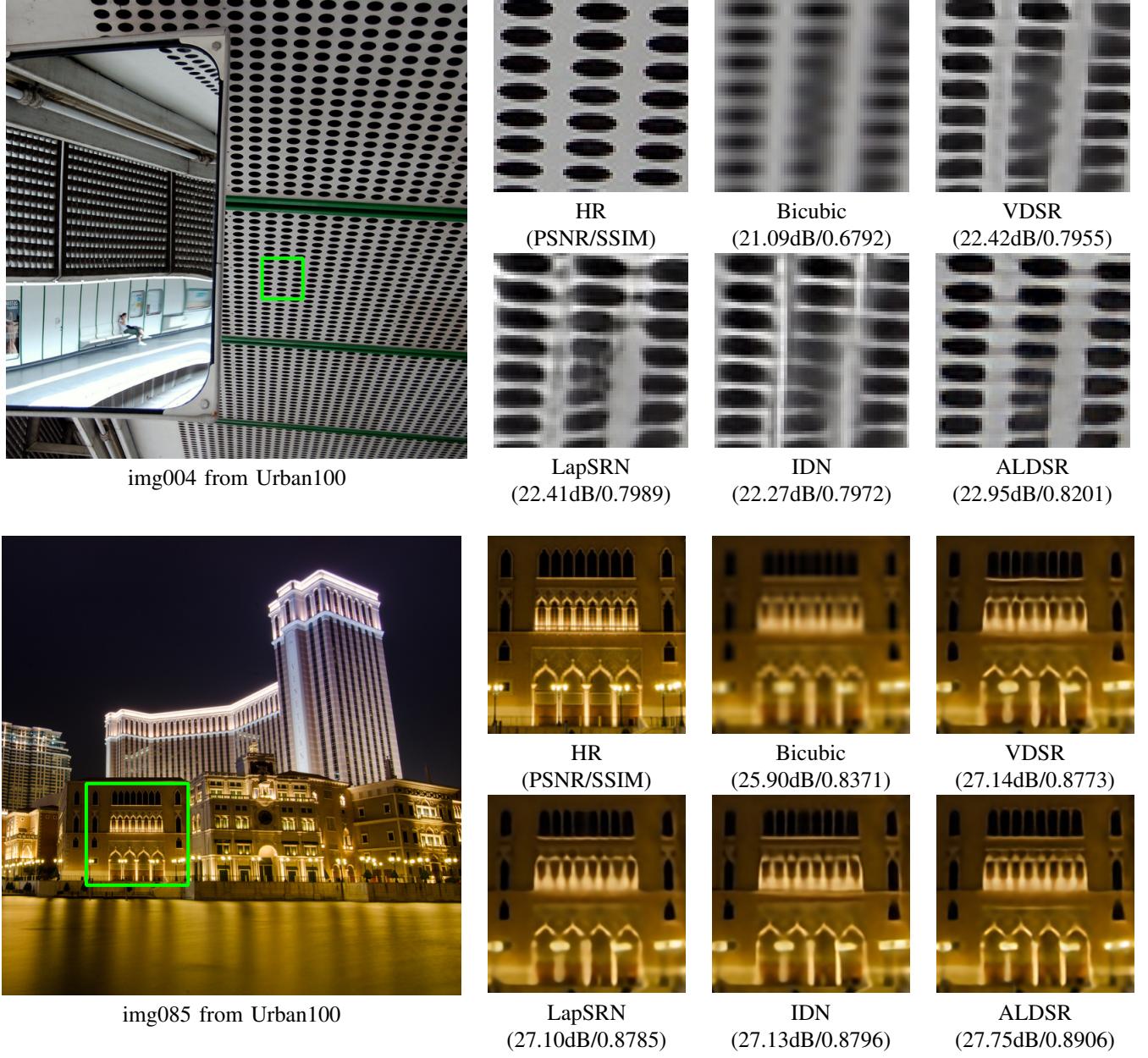


Fig. 7: Visual comparison for 4 $\times$  SR on Urban100 dataset.

- [3] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.
- [4] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.
- [5] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2472–2481.
- [6] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilennets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
- [7] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
- [8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [9] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 286–301.
- [10] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11065–11074.
- [11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [12] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.
- [13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [14] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop,

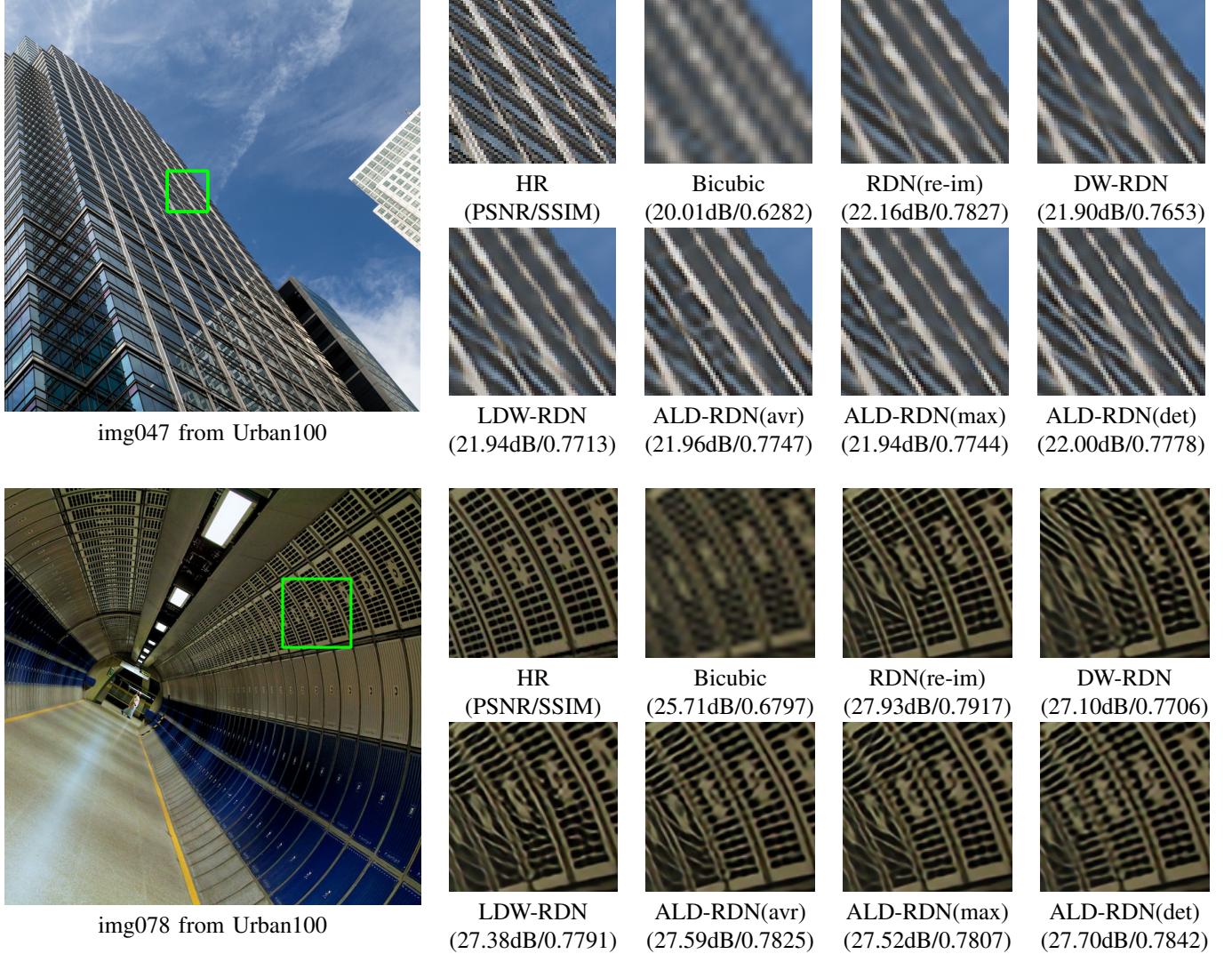


Fig. 8: Visual comparison for 4× SR on Urban100 dataset.

- D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.
- [15] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Deep laplacian pyramid networks for fast and accurate super-resolution,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 624–632.
- [16] Y. Tai, J. Yang, X. Liu, and C. Xu, “Memnet: A persistent memory network for image restoration,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4539–4547.
- [17] Z. Hui, X. Wang, and X. Gao, “Fast and accurate single image super-resolution via information distillation network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 723–731.
- [18] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, “Ntire 2017 challenge on single image super-resolution: Methods and results,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 114–125.
- [19] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, “Low-complexity single-image super-resolution based on nonnegative neighbor embedding,” 2012.
- [20] R. Zeyde, M. Elad, and M. Protter, “On single image scale-up using sparse-representations,” in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730.
- [21] D. Martin, C. Fowlkes, D. Tal, J. Malik *et al.*, “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics.” Iccv Vancouver, 2001.
- [22] J.-B. Huang, A. Singh, and N. Ahuja, “Single image super-resolution from transformed self-exemplars,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5197–5206.
- [23] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [24] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” 2017.



**Seongmin Hwang** received the B.S. and M.S. degrees in electronic engineering from Chonnam National University, South Korea, in 2017 and 2019, respectively. He is currently pursing the Ph.D. degree in the same university. His research interests include image processing, computer vision, machine learning and deep learning.



**Gwanghuyn Yu** received the B.S. degree in electronic engineering from Chosun University, South Korea in 2016. He got the M.S. degree in electronic engineering from Chonnam National University, South Korea in 2018. Since 2018, he has been a Ph.D. student in the same university. Also, he is a CEO from Insectpedia company, South Korea. His main research area includes digital signal processing, image processing, speech signal processing and machine learning.



**Cheolkon Jung** (M'08) received the B.S., M.S., and Ph.D. degrees in electronic engineering from Sungkyunkwan University, South Korea, in 1995, 1997, and 2002, respectively. He was a Research Staff Member with the Samsung Advanced Institute of Technology (Samsung Electronics), South Korea, from 2002 to 2007. He was a Research Professor with the School of Information and Communication Engineering, Sungkyunkwan University, from 2007 to 2009. Since 2009, he has been with the School of Electronic Engineering, Xidian University, China, where he is currently a Full Professor and the Director of the Xidian Media Lab. His main research interests include image and video processing, computer vision, pattern recognition, machine learning, computational photography, video coding, virtual reality, information fusion, multimedia content analysis and management, and 3DTV.



**Jinyong Kim** received the B.S., M.S., and Ph.D. degrees in electronic engineering from Seoul National University, South Korea, in 1986, 1988, and 1994, respectively. He was a full time researcher at Korea Telecom Software Research Center, South Korea, from 1993 to 1994. Since 1995, he has been a Full Professor with School of Electronics and Computer Engineering, Chonnam National University, South Korea. His main research topics include digital signal processing, image processing, speech signal processing and machine learning.