IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 August 2023, accepted 21 September 2023, date of publication 4 October 2023, date of current version 8 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3321976

METHODS

Dual-Stage Super-Resolution for Edge Devices

SAEM PARK[1,2], GEUNJAE CHOI[1,2], SEONGUK PARK[1], AND NOJUN KWAK[1], (Senior Member, IEEE)

[1]Graduate School of Convergence Science and Technology, Seoul National University, Gwanak-gu, Seoul 08826, Republic of Korea
[2]LG Electronics, Seocho-gu, Seoul 06772, Republic of Korea

Corresponding author: Nojun Kwak (nojunk@snu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) through Korean Government under Grant 2021R1A2C3006659, in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) through Korean Government under Grant 2021-0-01343, and in part by LG Electronics.

ABSTRACT To reduce memory usage, edge devices such as TVs use Super-resolution(SR) with dedicated hardware networks. Dedicated hardware has the disadvantage of being difficult to change and difficult to improve performance. We propose a dual-structured SR system that can extend this dedicated hardware, suggesting a way to improve the restoration performance for compression with minimal change in hardware. Edge devices such as general TVs deal with moving videos rather than still images. In this case, the video includes video compression, and a preprocessing step of restoring deterioration is useful. We propose as the main idea that these steps can be handled while removing only pixel shuffle layers from existing networks. Our proposed SR method consists of two stages: 1) restoring from the video compression without changing the size of LR images and 2) increasing the resolution. The first stage can be learned by reducing the difference between video-LR (Low-Resolution video images with codec degradation) and downscaled-HR (video images made without codec degradation by simply reducing the size of High-Resolution video). The second stage, resolution enhancement, performs the same task as the traditional SISR task, except that it focuses on restoring the output of the first stage rather than a downscaled-HR. Our new dataset for this processing, HD2UHD, consists of (video-LR, downscaled-HR, and HR) tuples. We also propose a new scheme of input distillation that utilizes video-LR and downscaled-HR at the same time.

INDEX TERMS Super-resolution, removal of video degradation, hardware reuse.

I. INTRODUCTION

Edge devices such as TV mainly deal with videos as input rather than still images. We want to find a direction to utilize the dedicated SR network due to the characteristics of the video. Therefore, in the introduction, the characteristics of the video and the dedicated SR will be described.

A. NECESSITY OF DEDICATED SR HARDWARE

In general, GPUs or NPUs are used to process deep learning. In this case, the results and weight parameters of each layer are stored in memory each time and are repeatedly called when necessary. In tasks that analyze the entire image, such as classification or segmentation, which can be processed by reducing the input size, it is efficient to use GPUs or NPUs like this. However, in the Super-resolution task, which improves the detail of the image, the size of the input cannot be reduced, and the image must be processed on the fly without frame delay. Due to this constraint, SRs on edge devices use dedicated CNN hardware rather than NPU or GPU. Nowadays, most broadcasts are composed of FHD or lower, and UHD TVs are popular, SR is one of the essential elements of TV, and the development of a dedicated SR network is necessary. The SR network developed once in this way has a disadvantage in that it is difficult to expand further, resulting in a problem to improve additional performance.

B. SINGLE IMAGE SUPER RESOLUTION (SISR)

In the conventional deep learning-based Super-resolution (SR) task, training datasets consist only of target images in a high resolution (HR), which are downscaled to low resolution (LR) images by a scaler method such as bicubic

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

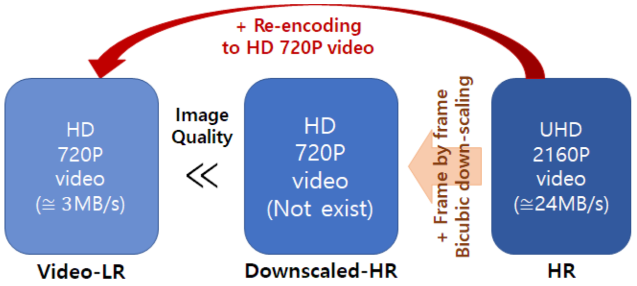


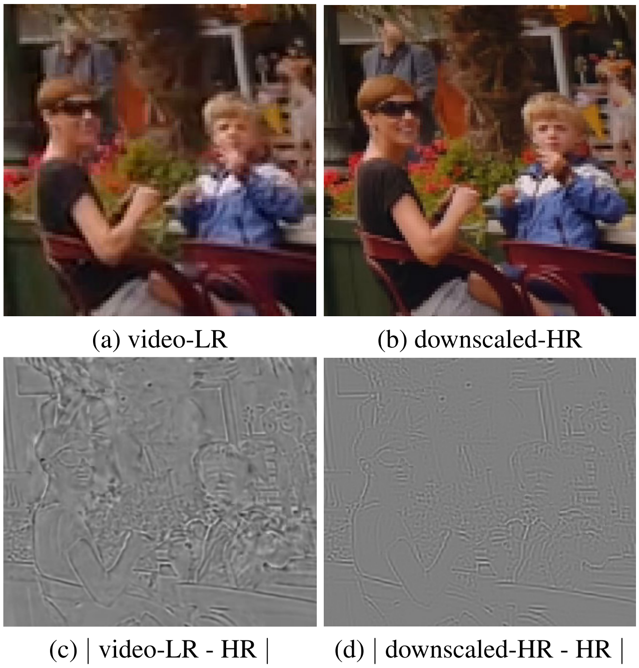
FIGURE 1. Definition of HR, downscaled-HR, and video-LR In this paper, we set a UHD (2160P) video as an HR, the target of learning. downscaled-HR is a video that is obtained by scaling downing HR on a frame-by-frame basis, which experiences only image-level degradation. Finally, video-LR refers to an actual HD (720P) video converted by a video codec encoding.

interpolation, and SR algorithms learn how to restore the target HR image from the LR image. Since SRCNN [1] successfully tackled this single-image-based SR (SISR) task using a deep neural network (DNN), a lot of algorithms have been developed with the growth of both network size and performance [2], [3], [4]. Although the conventional SISR algorithms restore the resolution degradation factor between an HR image and the corresponding LR image, which is a downscaled-HR, very well, they do not cover the gap between the downscaled-HR and the low-resolution video (video-LR) encoded by a video encoder such as MPEG4 or H.264 in the real world. We would like to add a one-step process to the SR developed in the form of SISR mentioned above and to explain this, we would like to separate degradation caused by simple resize from degradation caused by video compression.

C. DEFINITION OF HR, DOWNSCALED-HR, AND VIDEO-LR

To clarify a further discussion, as shown in Fig. 1, we define three types of images: HR, downscaled-HR, and video-LR.

1. HR is the target of SR which corresponds to the conventional dataset for SISR. In previous works, relatively low-resolution image datasets such as SET-5 [5], SET-14 [6], and Urban100 [7] were used as restoration goals. However, since we are aiming to develop an effective SR for real applications, we set to use UHD (2160P) video frames as HR images, the resolution of which is mainly supported in mainstream displays nowadays.
2. Downscaled-HR refers to images in which the resolution of HR is reduced using a downscaling such as bicubic interpolation. The conventional SISR algorithms use DNNs to learn to recover downscaled-HR to HR. downscaled-HR involves only image-level degradation from HR, in which high-frequency components that cannot be expressed at small resolutions are lost and aliasing is formed. Currently, what's commonly referred to as high-resolution images are FHD (1080P) or higher-resolution images, and we wanted to set up a task to restore the low-resolution video to a UHD level. Therefore, downsizing is processed by reducing a UHD (2160P) video by 1/3 to the HD (720P) level with the bicubic downsampler.


(a) video-LR (b) downscaled-HR

(c) | video-LR - HR | (d) | downscaled-HR - HR |

FIGURE 2. Cropped sample of video-LR and downscaled-HR. The video-LR in (a) and the downscaled-HR in (b) show significant quality differences despite having the same resolution. This can be confirmed more easily by looking at the difference between video-LR and HR in (c) and that between downscaled-HR and HR in (d). Although the difference in (c) is much larger than the difference in (d), conventional SR only learns the difference in (d).

3. Video-LR refers to an original HD (720P) video that is encoded in small size and can be directly downloaded or played back. This includes codec compression, so video-LR has much worse image quality than downscaled-HR. Learning a network to convert a video-LR (720P) into an HR (2160P) is very important and is the key to making SR a more meaningful technique. Therefore, in this paper, we tackle this SVISR problem and try to find a way to restore this video-level degradation. The quality of video-LR is much worse than the downscaled-HR as confirmed in Fig. 2, where (c) and (d) show the difference between the HR and the upscaled version of video-LR and downscaled-HR, respectively.

Rather than relying on conventional SISR settings (downscaled-HR → HR), we should tackle the new problem of Single Video Image SR(SVISR) (video-LR → HR) for more useful applications of SR techniques. We would like to propose a dual-stage SR with a step of reducing video degradation to improve the performance of the existing developed SR. At this time, it is intended to help develop edge devices by making it possible to use the previously developed SR network without change. In addition, adopting the scheme of knowledge distillation [8], we propose input distillation to restore lost texture components between video-LR and downscaled-HR by improving the existing

SISR which focuses only on the restoration of edge transitions (see Fig. 2(d)).

D. CONTRIBUTION

Our contributions can be summarized as follows:

1) We propose a system that can extend the capabilities of existing developed deep learning SRs while minimizing changes to the hardware. The added pre-stage is designed to be effective in reducing degradation due to video compression.
2) We propose a new input distillation method. This is a method of putting different inputs into the same network, the better of which is the teacher and the other is the student. When performing distillation with the result of stage 1 as the student and downscaled-HR as the teacher, it can achieve better results than training just with loss functions based only on the difference between the input LR-video and target HR images.

II. RELATED WORKS

A. BLIND SUPER-RESOLUTION

In contrast to the SISR, where degradations for inputs are specified and the inversion of bicubic interpolation is learned, there is a concept of blind Super-resolution or real-world Super-resolution to make SR effective in situations where degradations cannot be specified [9], [10], [11], [12]. Among them, Real-ESRGAN [13], which added the concept of blind SR to ESRGAN [14], is a representative work.

Real-ESRGAN raised the problem of the conventional ESR-GAN that generates an LR image by bicubic interpolation and suggested that LR images should be reconstructed by considering degradation such as jpeg compression. It can be seen as a work dealing with a similar problem to our paper. However, even the Real-ESRGAN relies on still image-based degradation of an HR image, so only image-level degradation is learned, and the actual video-level degradation, which is much more serious, is not considered, resulting in low-quality outputs for compressed videos. Unlike the Real-ESRGAN which learns about arbitrary virtual degradation, converting various degradations by a single model and changing the input image into an animation-like image with sharp boundaries, we have obtained more natural results through real video-based learning.

B. VIDEO SUPER-RESOLUTION

Unlike SISR which receives only a still image as input, video SR is a task aimed at improving the quality of videos. Video SRs that use multi-frame methods are specialized for sequential video and are trained to receive multiple consecutive frames as inputs and produce a single image [15], [16], [17]. There are other video SRs such as VideoSR-CNN [16], Deep VideoSR [18], basicVSR [19]. And video-based data such as VID4 [20], REDS [21] and VIMEO90K [22] have been proposed for video learning. Although these algorithms are designed to receive consecutive video frames as inputs and produce better results from increased input channels, the increased network costs and memory buffers are the weaknesses. In addition, continuous video is required, which requires additional heuristics, such as scene switching or initial reset. Thus, we intend to create a solution suitable for video processing without using such a multi-frame approach, while maintaining a typical single-input single-output SR system simply by dual-stage training of the SR network on a dataset with improved quality. And most importantly, the task of video SR is based on single-video impaired by a scaler, just like single-image-based SR, not a concept to learn degradation due to a video codec.

III. METHOD

A. DUAL-STAGE SR SYSTEM

As mentioned in the introduction, we note that conventional SRs can only learn image-level degradation due to self-image transformation, which has a very large gap from videos in the real world. So, reducing this gap is very important especially when real-time SR processing of the video is needed so that we cannot apply multi-frame SR for videos. In Edge devices, multi-frames are difficult to use for SRs due to the increased use of frame-storing memory buffers and the occurrence of side effects on scene transitions, as previously mentioned in the introduction. Since conventional methods did not consider video-level degradation, we added an additional network to the existing SR network for restoration of video-level degradation and applied it to reduce the difference between real low-quality videos and high-quality videos with only simple down-scaling. Fig 3 shows the conceptual diagram of our proposal. We propose a dual-stage SR, responsible for restoring video compression in the first stage and enhancing the resolution in the second stage.

This dual-stage SR can be represented as follows:

$$\hat{I}_{HR} = Net_2(\hat{I}_{dHR}|\theta_2),$$
$$\text{where } \hat{I}_{dHR} = Net_1(I_{LR}|\theta_1). \tag{1}$$

Each stage uses its respective L_1 loss as follows:

$$\mathcal{L}_1 = \|Net_1(I_{LR}|\theta_1) - I_{dHR}\|_1$$
$$\mathcal{L}_2 = \|Net_2(\hat{I}_{dHR}|\theta_2) - I_{HR}\|_1. \tag{2}$$

Here, I_{LR}, I_{dHR}, and I_{HR} represent a video-LR, a downscaled-HR, and an HR image respectively. $Net_1(\theta_1)$ and $Net_2(\theta_2)$ are the SR network (parameter) for the first and the second stages. The hat notation ($\hat{I}_{dHR}$, $\hat{I}_{HR}$) is used for indicating the estimated image. Then, the parameters for the two networks are trained in an end-to-end manner by

$$(\theta_1^*, \theta_2^*) = \underset{(\theta_1,\theta_2)}{\arg\min}\, \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2, \tag{3}$$

where we use $\lambda_1 = \lambda_2$ for simplicity.

From the network point of view, the difference between the first stage and the second stage can be seen as the presence or absence of the final pixel-shuffle layer. In the

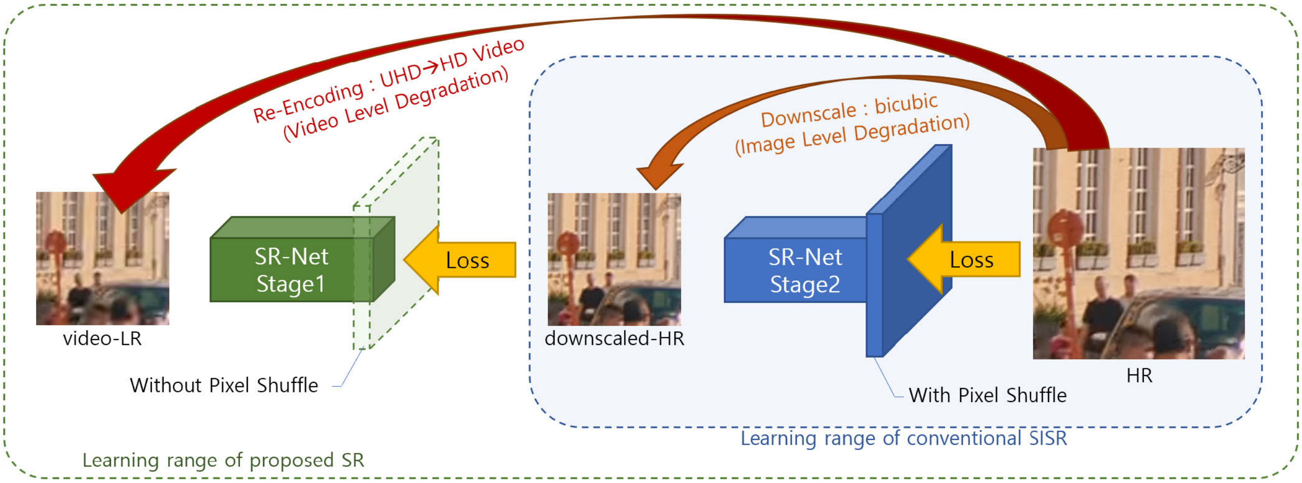

Re-Encoding : UHD→HD Video
(Video Level Degradation)
Downscale : bicubic
(Image Level Degradation)
SR-Net Stage1
Loss
SR-Net Stage2
Loss
video-LR
downscaled-HR
HR
Without Pixel Shuffle
With Pixel Shuffle
Learning range of conventional SISR
Learning range of proposed SR

FIGURE 3. The concept of our Dual-stage SR. The key idea is to put the conventional SISR, which focuses on the restoration of the image-level degradation (*e.g.*by bicubic interpolation), as stage 2 and to have an additional convolution for restoring the video-level degradation (*e.g.*by video codec) at stage 1 by inputting a real video-LR image. The difference between the first and the second stages lies in the presence or absence of the pixel shuffling operation to change the size of the output. The proposed dual-SR system trained by the proposed data set, HD2UHD, made of real videos provides substantial improvement in the low-quality videos.

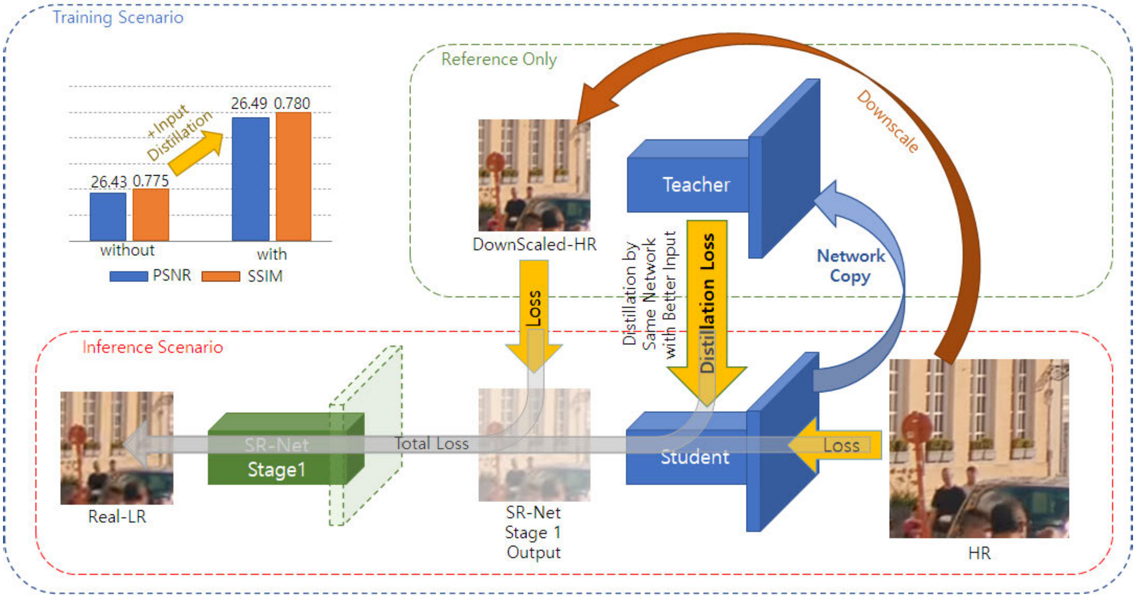


FIGURE 4. Overview of the proposed dual-stage SR system including input distillation. This figure shows the concept of 'input distillation' and how it works. For our dual-SR system, distillation loss is added to the end-to-end loss of each stage and affects the entire learning procedure.

first stage, the network learns to map video-LR inputs of the same size to downscaled-HR without a pixel-shuffle layer, and in the second stage, it learns to map to a larger-sized HR through a pixel-shuffle. As a result, SR for the first stage can be used without requiring independent development, removing only the last layer from the existing HW.

For these two stages of learning, we publish a new training dataset, HD2UHD, presented in Sec. III-B each tuple of which consists of separately downloaded LR and HR and an additional intermediate downscaled-HR image.

B. HD2UHD DATASET

Recently, many applications have emerged that transmit the same content at various bit rates according to the network and viewing situations. Typical examples include Netflix and YouTube, where content providers offer the same original video in various formats, including SD, HD, FHD, UHD, and 8K videos.

By using this, the following three versions of the same content can be obtained. 1) Receive a UHD version and set it as HR, the target image. 2) Create a downscaled-HR that acts as an intermediate target by scaling down the HR image frame-by-frame. 3) Download the HD version and use it as a video-LR input.

We collected various genres of Internet videos such as natural scenery, cityscape, music video, sports, animation, games, and CG, and extracted 10 random patches from each frame to create a total of more than 660,000 patches. For the flexibility of future learning, each patch of video-LR was cropped into 128×128, which is spatially larger than other datasets typically in 64 × 64, and HR was cropped into 384×384, whose both horizontal and vertical sizes are three times of LR. The downscaled-HR reduced the resolution of HR to 128 × 128 using a bicubic scaler for the HR data set stored in a still-image state. In this paper, we analyze and learn based on this dataset, and also try to verify the results by creating 400 test sets with new videos that are not used for training in the same way. Our new dataset is named HD2UHD, which will be publicly available with this paper.

C. INPUT DISTILLATION

An interesting point of our proposed system is that the internal target and the intermediate processing result have the same size. In other words, the second stage can operate with both the result of the first stage and the downscaled-HR as an input. In general, knowledge distillation [8] is aimed at obtaining better performance by using features made from better networks as teachers. Downscaled-HR is not used in the case of inference but only affects learning.

We focused on the fact that we could apply distillation to something that works better based on the input rather than a network. Since downscaled-HR is a more correct answer that is much more similar to the HR image than the processed video-LR result, even if the two networks have identical parameters, it can be seen that when downscaled-HR is processed as an input, it produces better features than when video-LR is used.

In each iteration, the second stage updated with the latest weight is operated twice 1) with the normal case (using stage 1's output as input) and 2) downscaled-HR as input:

$$\begin{aligned}\hat{I}_{HR,s} &= Net_2(\hat{I}_{dHR}|\theta_2)\\ \hat{I}_{HR,t} &= Net_2(I_{dHR}|\theta_2)\\ \text{where } \hat{I}_{dHR} &= Net_1(I_{LR}|\theta_1).\end{aligned} \tag{4}$$

Then, the case where downscaled-HR is used is taken as a teacher and the remaining normal case is taken as a student for feature map-level knowledge distillation [23]. It assumes that a student who solves a problem by watching the internal answer can teach a better solution to a student who solves the problem from the beginning. This is a novel method of applying distillation using different inputs on the same network, unlike the usual method of applying distillation between different networks. We refer to this technique as 'input distillation', which distills features of a better input on a network to features of a worse input on the same network. The input distillation loss can be defined as follows:

$$\mathcal{L}_{id} = \sum_{l\in[L]} \|F^l_{Net_2}(\hat{I}_{dHR}|\theta_2) - F^l_{Net_2}(I_{dHR}|\theta_2)\|_1, \tag{5}$$

where L denotes the total number of layers in the second network (Net_2) and F^l denotes the feature map of the l-th layer.

As in Eq. (5), the distillation loss consists of the sum of L1 differences between each feature map of internal layers. This concept of input distillation can be easily understood in Fig. 4. It includes all of our proposed dual-stage SR setup, including input distillation. The second stage in our dual-stage SR is performed once more with the downscaled-HR in addition to the basic operation, and input distillation is applied. Note that the distillation loss is only propagated to the student ($F^l_{Net_2}(\hat{I}_{dHR}|\theta_2)$) and the network parameters (θ_2) are copied from the student to the teacher at each iteration as shown in Fig. 4.

End-to-end losses of each step, $\mathcal{L}_1$ and $\mathcal{L}_2$ in (2), are combined with input distillation loss and backpropagated:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2 + \lambda_{id}\mathcal{L}_{id}. \tag{6}$$

The distillation in the feature domain has the advantage of being capable of non-linear conversion on both sides of several tens of channels [24], while the information on the final output image is only 3 channels and one side is fixed. Moreover, the output end of the feature is fixed to converge to the HR image, resulting in stable information without divergence. More information can be extracted through this expanded channel, resulting in better results when applying input distillation. Distilling the feature maps instead of the final output images is well-verified in several previous works on Super-resolution [25], [26], [27].

The great advantage of the application of the proposed distillation is that the performance can be improved with the same HW by changing only the loss to learning. As mentioned in the introduction, we tried to explore ways to improve performance using the previously developed dedicated SR network for edge devices. As a solution to this, we present a dual-stage SR system that can further remove video degradation, and propose distillation loss that can learn this degradation more effectively.

IV. EXPERIMENTAL RESULTS

A. EXPERIMENT SETTINGS

Within each of the two stages, it is possible to utilize a latest SR network and loss settings as they are, so we would like to apply the proposal to various existing SR methods to confirm the effectiveness of our proposal. We tried to apply our proposals on top of existing SRs ranging from SRCNN, which is the most similar to the dedicated HW, to recent RCAN. The basic learning system of each SR was set up as the second stage of the proposed dual-stage SR, while for the first stage, the scaling factor was changed to 1 or only the pixel-shuffle was removed. Since there are no restrictions at each stage and each of our proposals (HD2UHD dataset, Dual-SR, input distillation) can be dealt with separately, we checked the effect of our algorithms using various SRs.

For the experiments, we set up a simple environment based on the baseline architecture of RCAN [28] using L_1 loss only.

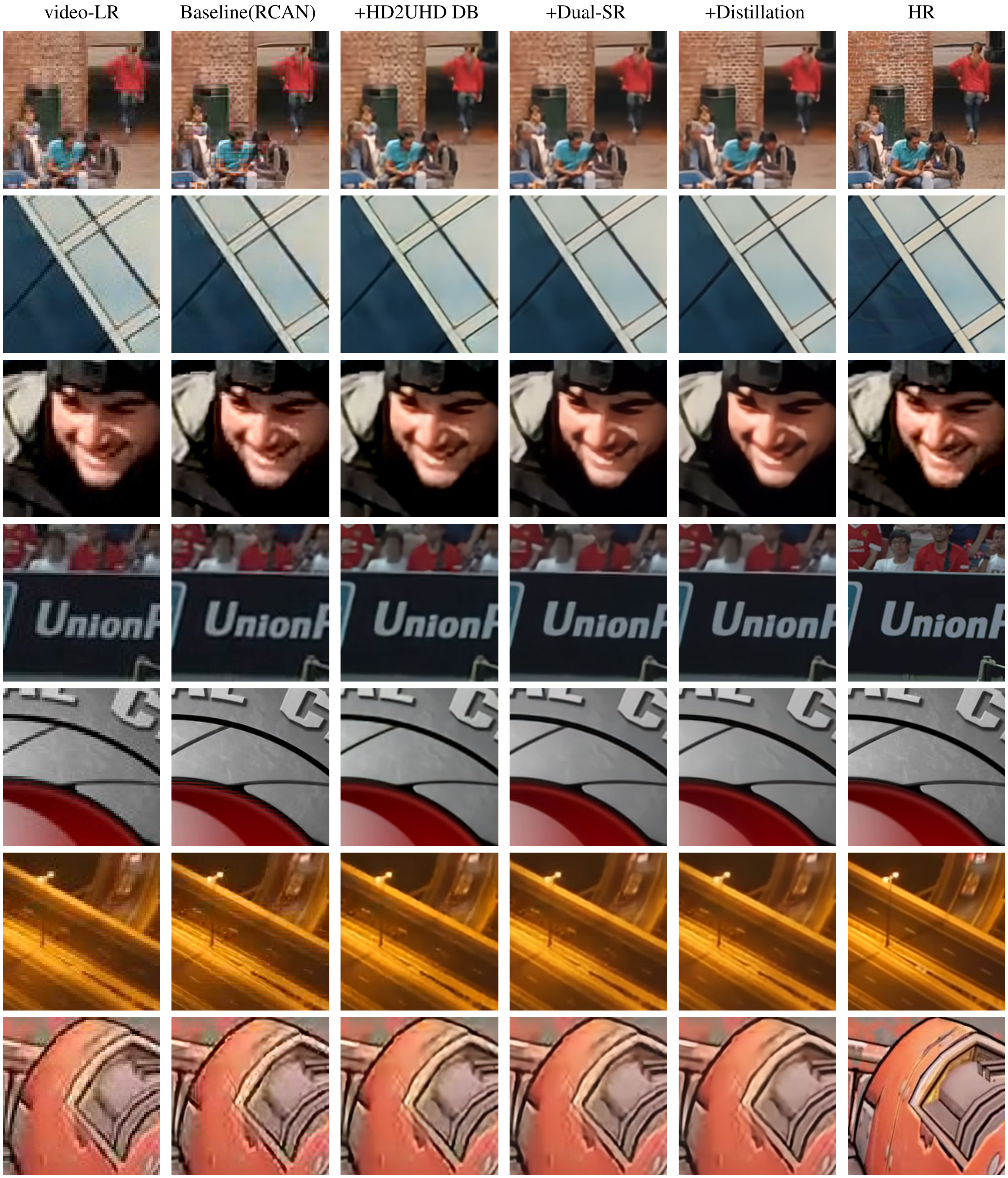


FIGURE 5. Experimental results. These images show what changes occur as the proposed methods are applied one by one to the RCAN baseline. Each column (excluding HR) is a result of applying an additional proposed solution to its left column. While the baseline incorrectly boosts the grid pattern due to the video compression, the result of learning with our proposed HD2UHD dataset does not generate such artifacts. Dual SR shows better results even though both stages of which use a half-sized network compared to the baseline to make the total number of parameters unchanged. Additionally, when input distillation is applied, it can be seen that the shape of the texture becomes closer to that of HR.

When applying GAN loss [29] or perceptual loss [30], the problem of reproduction arises and the effectiveness of the proposed concept faded, so we set up without these losses to confirm the effectiveness of the proposed method clearly.

The RCAN network uses 10 residual groups each consisting of 6 residual blocks, and in the case of Dual-SR, the number of residual groups was reduced by half to limit the total number of parameters. In our setup, each network is only a

TABLE 1. Quantitative evaluation. In the real HD videos, all three of our proposed methods have improved scores.

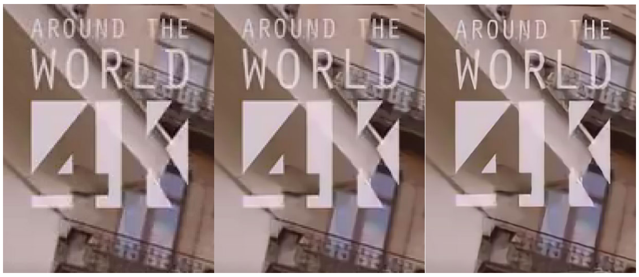
Network	Training Method	PSNR	SSIM
SRCNN	Baseline (dHR→HR)	25.38	0.7636
	+HD2UHD Dataset	25.64	0.7644
	+input distillation	25.68	0.7665
RCAN	Baseline (dHR→HR)	26.24	0.7675
	+HD2UHD Dataset	26.36	0.7714
	+Dual-SR	26.43	0.7752
	+input distillation	26.49	0.7797

small SR with 5M parameters, and Dual-SR with two stages is designed to have a total of 10M parameters. This is a lighter SR than the 15M RCAN baseline.

B. QUALITATIVE RESULTS

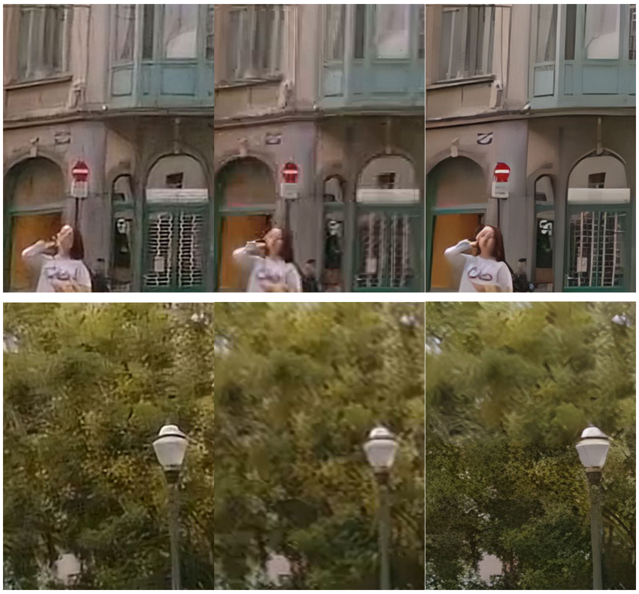
The qualitative experiment shows the change in the image by gradually adding each component of our proposed method to the RCAN baseline trained on bicubic degradation. As introduced in Sec. III, there are three major proposals in this paper: 1. HD2UHD dataset, 2. Dual-SR architecture and 3. input distillation.

The main qualitative results are shown on Fig. 5. The first column magnifies video-LRs encoded with HD (720P) as the input data, and the last column is original UHD (2160P) video images used as the learning target. The second column is the result of RCAN-baseline for video-LR inputs. When video-LR is used as input, the results clearly exhibit flaws, especially at the edges. It can be seen that lattices are forming around the chroma components as well as near strong edges. The third column is the result of changing only the training dataset to our HD2UHD dataset while maintaining the baseline method of RCAN's network and training method. In this case, it can be seen that the wrong grid disappears and the image as a whole becomes smoother at the cost of generating more blurry images. The fourth column is the result of applying the proposed Dual-SR learning system together with the change of the training dataset. Although the network was divided into two stages and maintained the same level of parameters, the results show clearer and smoother improvements. Similar tendencies additionally occur when input distillation is applied. In the first image, it can be seen that parts such as the texture of the brick resemble HR more stably, and the softness and sharpness of the edge are also improved. We showed that each of our three proposals has effects on improving image quality towards that of the original HR.



video-LR 1st-Stage Output 2nd-Stage Output

FIGURE 6. Results in the actual edge device. These images are the result of storing the actual HW output using Alpha 9 Gen-6 DTV SoC. The noise caused by compression was removed in the first stage, and the Super-resolution was well performed in the second stage.



Ours VSR++ RealBasicVSR

FIGURE 7. Comparison with VSR++ and RealBasicVSR. The result of an HD video input for comparison with VSR++ and RealBasicVSR. Unlike basicVSR, which over-boosts certain textures or creates artifacts, the proposed dual-SR shows stable results across all videos.

C. RESULTS IN THE ACTUAL EDGE DEVICE

We checked the performance of these dual SR systems based on the actual dedicated HW. This dual video quality improvement processing is mounted on *LG* α9 Gen-6, contributing to image quality improvement. We constructed the same learning system using a quantized, small dedicated SR network at a level where real-time processing can be implemented and we want to show the learning results using it. Fig 6 is this result. Although we cannot disclose this network and quantitative evaluation results due to technology security, we would like to show that the proposed dual-SR system and input distortion as a result of image improvement helped develop HW for actual edge devices. We have actually applied the proposed system to the Super-resolution algorithm in TV SoC to improve performance without requiring additional development time by reusing the existing HW.

To ensure that Dual-SR algorithm is competitive with state-of-the-art video SRs, we also compared it to VSR++ and RealBasicVSR in Fig 7. Applying basicVSR to edge devices is very challenging because it adds a huge computation of extracting separate optical flows and synthesizing the latent vectors of several frames at the same location. Nevertheless, VSR++, trained to remove only certain artifacts well, produced unusable results on random HD video inputs with severe image jaggies like RCAN's baseline model.

TABLE 2. **Quantitative analysis of the effect of distillation loss. When the second network receives a downscaled-HR image as an input rather than the output of the first stage, the effect of input distillation is prominent, improving both PSNR&SSIM.**

Network	Training Method	Input	PSNR	SSIM
Net_2	Dual-SR+L_1only	dHR	26.16	0.7259
	Dual-SR+Distillation		30.49	0.8440

While RealBasicVSR, which is trained to restore arbitrary degradation well, is relatively stable and has a good level of cleanliness and enhancement, it has been shown to over-boost certain textures unevenly. The proposed Dual-SR has the advantage of a cheap network and easy computation, so the total amount of enhancement is smaller than basicVSR, but it works stably in all situations.

D. QUANTITATIVE EVALUATION

For quantitative evaluation, we used a total of 400 patches extracted from additional new videos to measure PSNR and SSIM [31]. The validation set includes dozens of videos from various genres, including animation, natural scenery, urban scenery, CG, movies, and music videos. Following the convention, PSNR and SSIM were measured using the difference in luma (Y) after YCbCr conversion.

The experimental results are shown in Table 1. The tendency seen in the qualitative results could be objectively confirmed. The proposed HD2UHD validation set, which mainly deals with HD (720P) resolution videos, showed additional improvement effects compared to the baseline, and particularly, a large difference could be confirmed when using SRCNN. In the case of RCAN, the score difference was not as dramatic as the difference in Fig. 5, but the overall direction of improvement could be confirmed. We can see that Dual-SR and input distillation have an additional improvement effect when applied one by one.

To confirm the effectiveness of the input distillation more objectively, we checked how closer becomes the resultant output to the target HR with the application of input distillation when the second network receives a downscaled-HR image as an input. Table 2 confirms that Net_2's teacher output has become similar to HR with input distillation, so we can argue that the distribution of $Net_1(video\text{-}LR)$ has become similar to downscaled-HR.

V. CONCLUSION

We propose dual-SR, which learns existing SRs in two stages for learning, and an input distillation method that can more actively utilize downscaled-HR as an intermediate answer. We confirmed the effect by applying each of the proposed HD2UHD databases, dual-SR, and input distillation to the existing RCAN and SRCNN baselines. The effectiveness of the proposed method is demonstrated by using quantitative and quantitative evaluation and it is shown to eliminate side effects of conventional methods in low-quality videos and obtain more suitable final SR results.

For future works, we want to do ablation studies with other SRs that use heavier networks and various losses than RCAN. We believe that dual-SR will be beneficial for SoTA-class SRs as well, and expect to see good results in various videos.

REFERENCES

[1] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[2] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1132–1140.

[3] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[4] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[5] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. British Mach. Vis. Conf.* BMVA Press, 2012, pp. 135.1–135.10.

[6] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surf.* Cham, Switzerland: Springer, 2010, pp. 711–730.

[7] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5197–5206.

[8] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[9] T. Michaeli and M. Irani, "Nonparametric blind super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 945–952.

[10] X. Ji, Y. Cao, Y. Tai, C. Wang, J. Li, and F. Huang, "Real-world super-resolution via kernel estimation and noise injection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1914–1923.

[11] J. Gu, H. Lu, W. Zuo, and C. Dong, "Blind super-resolution with iterative kernel correction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1604–1613.

[12] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-gan," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst., Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 32, Dec. 2019, Pages 284–293.

[13] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1905–1914.

[14] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 1–10.

[15] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1327–1344, Oct. 2004.

[16] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Trans. Comput. Imag.*, vol. 2, no. 2, pp. 109–122, Jun. 2016.

[17] M. S. M. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-recurrent video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6626–6634.

[18] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct 2017, pp. 4472–4480.

[19] K. C. K. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "BasicVSR: The search for essential components in video super-resolution and beyond," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4945–4954.

[20] C. Liu and D. Sun, "A Bayesian approach to adaptive video super resolution," in *Proc. CVPR*, Jun. 2011, pp. 209–216.

[21] S. Nah, S. Baik, S. Hong, G. Moon, S. Son, R. Timofte, and K. M. Lee, "NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1996–2005.

[22] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *Int. J. Comput. Vis.*, vol. 127, no. 8, pp. 1106–1125, Aug. 2019.

[23] M. Ji, S. Shin, S. Hwang, G. Park, and I.-C. Moon, "Refine myself by teaching myself: Feature refinement via self-knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10659–10668.

[24] M. Z. Chen and J. Ming Wu, "Group feature information distillation network for single image super-resolution," in *Proc. 7th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2021, pp. 1827–1831.

[25] Z. He, T. Dai, J. Lu, Y. Jiang, and S.-T. Xia, "FAKD: Feature-affinity based knowledge distillation for efficient image super-resolution," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 518–522.

[26] S. Park and N. Kwak, "Local-selective feature distillation for single image super-resolution," 2021, *arXiv:2111.10988*.

[27] W. Lee, J. Lee, D. Kim, and B. Ham, "Learning with privileged information for efficient image super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 465–482.

[28] Z. Lin, P. Garg, A. Banerjee, S. Abdel Magid, D. Sun, Y. Zhang, L. Van Gool, D. Wei, and H. Pfister, "Revisiting RCAN: Improved training for image super-resolution," 2022, *arXiv:2201.11279*.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[30] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 694–711.

[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.

SAEM PARK was born in South Korea, in 1982. He received the B.S. and M.S. degrees in image science from the University of Chiba, Japan, in 2006 and 2008, with the Korea–Japan Young Engineers and Management of Government Scholarship. He is currently pursuing the Ph.D. degree with the Graduate School of Convergence Science and Technology, Seoul National University, South Korea. From 2008 to 2018, he was a Researcher with LG Electronics. He developed an image processing engine for TV Main SoC. From 2008 to 2015, he developed motion-compensated temporal noise reduction, de-jagging, motion adaptive deinterlacer, and de-blocking algorithms, for the LG XD Engine. In 2016, he developed edge shift super-resolution and de-contour algorithms for the LG Alpha 7 Engine. In 2017, he developed a sharpness enhancer, OLED logo sticky reduction, and object detection algorithms, for the LG Alpha 9 Engine. In 2022, he developed super-resolution for the LG Alpha 10 Engine.

GEUNJAE CHOI was born in Seoul, South Korea, in 1987. He received the B.S. and M.S. degrees in electrical engineering and computer science from Seoul National University, in 2009 and 2011, respectively, where he is currently pursuing the Ph.D. degree, with a focus on deep neural network optimization and lightweight design. Since 2014, he has been with LG Electronics, where he has been involved in the design of mobile application processors and deep neural network accelerator hardware.

SEONGUK PARK received the B.S. degree in electrical engineering from Sungkyunkwan University, Seoul, South Korea, in 2017. He is currently pursuing the Ph.D. degree with the Department of Intelligence and Information, Seoul National University, Seoul. His current research interests include generative models, model compression, and low-level vision.

NOJUN KWAK (Senior Member, IEEE) was born in Seoul, South Korea, in 1974. He received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from Seoul National University, Seoul, in 1997, 1999, and 2003, respectively. From 2003 to 2006, he was with Samsung Electronics, Seoul. In 2006, he joined Seoul National University as a BK21 Assistant Professor. From 2007 to 2013, he was a Faculty Member of the Department of Electrical and Computer Engineering, Ajou University, Suwon, South Korea. Since 2013, he has been with the Graduate School of Convergence Science and Technology, Seoul National University, where he is currently a Professor. His current research interests include feature learning by deep neural networks and their applications in various areas of pattern recognition, computer vision, and image processing.