Introduction to SUPERNOVA: A deep learning-based image/video quality enhancement platform

Taeyoung Na, Jin Jeon, Munkyeong Hwang, Juhan Bae, Jaeil Kim and

Jeongyeon Lim

SK telecom, Republic of Korea

ABSTRACT

Recently, various types of media services have drawn much attention with technical advances in the media processing arena. Numbers of IPTV/OTT based media services are becoming available through the Internet. This also becomes possible due to a stable installation of 5-G/LTE/3-G mobile network in addition to broadband networks. Thus, it is noted that the accessibility to media content increases and the demand for consuming high-quality media content also increases. Unfortunately, however, there still exists a lot of low-quality media content that needs to be enhanced. In this paper, we introduce a solution called SUPERNOVA that consists of deep learning-based methods to drastically enhance the quality of this media. Media content can be delivered to the SUPERNOVA platform through an API or more than one method can be selectively implemented in current local machines with GPUs. The current SUPERNOVA platform contains up-scaling (a.k.a super-resolution), HFR (High Frame Rate) and retargeting functions. It is noticed that both objective and subjective performance is clearly enhanced after applying each method in SUPERNOVA.

Section I: INTRODUCTION

With the rapid increase in the demand for image/video-based media services, quality of media content is becoming a more important topic. As is well known, image/video quality degradation is mainly due to the quantization during lossy coding process. This degradation becomes especially worse as customers are located where the transmission bandwidth becomes narrower because the bitrate for the encoded media contents' bitstream becomes lower in this environment. Another degradation case is when the spatial resolution for the delivered image/video is too small for customers to watch with their FHD or 4-K display. When this resolution degradation occurs due to instantaneous bandwidth constraints, the image/video will soon regain its original resolution, but the resolution degradation continues if the whole content in a CDN (Contents Delivery Network) or H/E Server are only stored with low resolution or low bitrate.

Until the early 2000s, most CPs (Contents Providers) produced their video contents with SD (720x480) resolution but resolutions of 4-K (3840x2160) beyond FHD (1920x1080) are currently supported in many mobile devices as well as TVs. In this case, viewers are exposed to fundamental deterioration in visual quality. Especially, the aspect ratios between SD and FHD/4K are 4:3 and 16:9, respectively. Thus, it requires more consideration when applying a linear up-scaling method such as super-resolution in order to maintain the shape of the original content. For FHD to 4-K UHD up-scaling case, 60fps is required when rendering 4-K content but most FHD content is only 30fps. In this case, methods that expand the frame rate need to be considered.

For all quality enhancement methods mentioned above, deep learning-based approaches are now considered. From the mid-2010s, deep leaning-based methods have been applied to both computer vision and media processing areas. These areas include the traditional processing area such as de-noising, up-scaling, re-targeting, SDR to HDR conversion, restoration and colorization as well as vision areas such as detection, recognition, etc. It is noticeable that all performance gaps before and after applying deep learning-based methods become wide although it requires heavy GPU computing powers. It is expected that the complexity of the deep learning network would increase because GPU cost is gradually decreasing. Figure 1 shows a conceptual process of image up-scaling in a deep neural network.

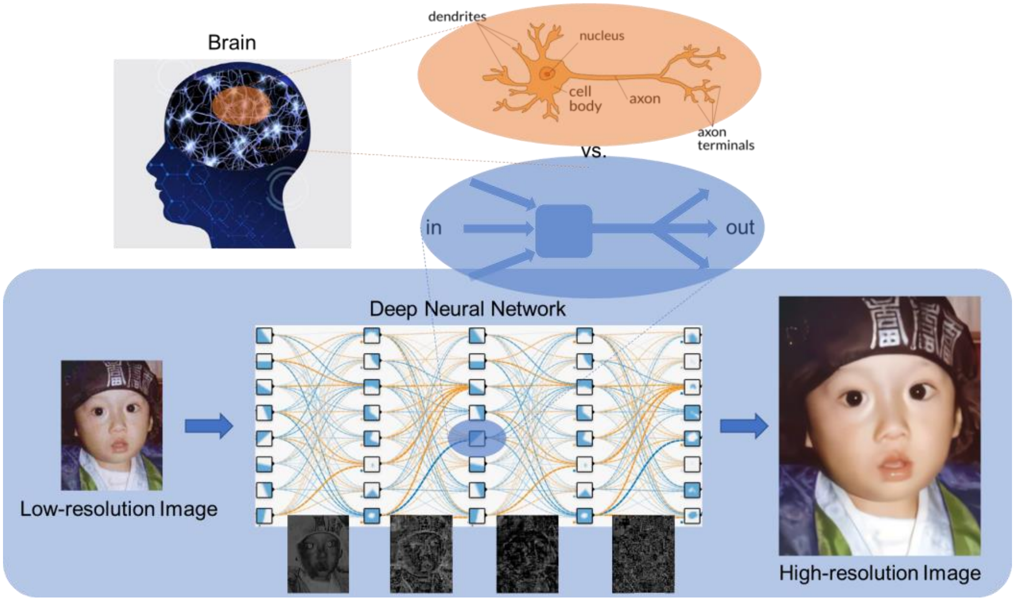


Figure 1. A conceptual process for image up-scaling

In this paper, deep learning-based media processing methods to enhance the visual quality of image/video contents in SUPERNOVA are presented. Then, we will show how these methods are combined in the SUPERNOVA platform and how to activate the SUPERNOVA functions.

The rest of the paper is organized as follows: In Section II, related work on media quality enhancement methods relevant to SUPERNOVA are reviewed; in Section III, the proposed methods in SUPERNOVA are presented; in Section IV, experimental results and performance comparisons are shown to verify the effectiveness of the proposed method; finally, we conclude our work in Section V.

Section II: RELATED WORK

Much research on enhancing the quality of media content has been discussed. As mentioned at Section I, deep learning-based work related to SUPERNOVA is mainly focused on, super-resolution, HFR and face image restoration, respectively. In particular, we give much attention to the state-of-the-art methods that achieve notable improvements.

Up-scaling and HFR: In contrast with conventional ResNet [1] and SRResNet [2] architecture, the enhanced deep super-resolution (EDSR) network proposed by Lim et al. [3] is optimized by eliminating unnecessary modules such as batch normalization to simplify the network architecture and employs residual scaling techniques to stably train large models, which results in reducing the heavy computation time and memory. In [4], a very deep residual channel attention network (RCAN) that adopts novel residual in residual (RIR) structures with several residual groups of long skip connections and each residual group consists of several residual blocks with short skip connections. A channel attention scheme is additionally proposed to adaptively rescale channel wise features by considering interdependencies among feature channels to concentrate on more useful channels. For HFR methods, Niklaus et al. [5] recently extended the 2-D kernel algorithm by estimating separable 1D kernels, which has greatly reduced the amount of required memory. In addition, Liu et al. [6] proposed a CNN-based deep voxel flow to produce dense voxel flow to optimize frame interpolation.

Image/Video Re-targeting: With the advent of new display technologies with variable form factors, media retargeting technology has been drawing much attention both in academia and industry. It is known that smartphone usage time in portrait mode is much more than that in landscape mode, but most image/video contents are produced and oriented in landscape mode. Thus, several media service providers attempt to produce content fit for portrait-view or convert original contents fit for landscape mode into contents fit for portrait-view. Michael et al. propose a method based on bi-directional warping (BDW) and they define a new image similarity measure. Furthermore, a dynamic programming algorithm to find an optimal path in the resizing space is proposed [7].

Face Image Restoration: Florian et al. propose an end-to-end training scheme and a novel network structure called FaceNet [8] for face verification, recognition and clustering. This method learns a mapping between face images and a compact Euclidean space where distances directly correspond to a measure of face similarity. In [9], a valuable dataset of human faces (Flickr-Faces-HQ, FFHQ) is provided. This dataset consists of 70M images of 1024x1024 resolution and can be useful in testing for face restoration work.

Section III: A PROPOSED IMAGE/VIDEO QUALITY ENHANCEMENT PLATFORM: SUPERNOVA

Figure 2. shows an architecture of the proposed SUPERNOVA that includes several processing modules.

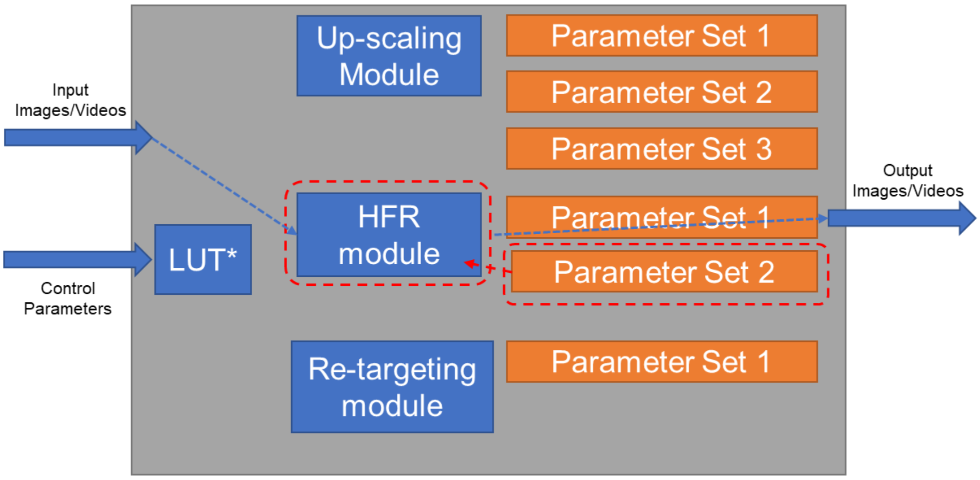


Figure 2. An architecture of the proposed SUPERNOVA

In Figure 2, the 3 distinct methods in SUPERNOVA are shown, which provide up-scaling, HFR and re-targeting functions, respectively. Obviously, each method has deep neural network with upgradable hyper parameter sets depicted as Parameter Set n in Figure 2. However, each module can be serially aligned to achieve better performance by Control Parameters as depicted in Figure 2. For example, one can call the HFR module before the up-scaling module or the re-targeting module. Although the image/video is degraded by multiple factors, SUPERNOVA provides users with maximal performance in this way.

Obviously, access to SUPERNOVA can be implemented through an API. Details of each module are covered in the following subsections.

1. Up-scaling (super-resolution) module

Deep learning-based super-resolution (SR) methods are actively being developed, and the performance has been significantly improved compared to previous methods before applying deep learning approaches. Although these SR methods achieve better performance, it is common that noticeable performance is not found when applying the present SR study on actual content for media services. This is because most dataset used in the present SR study are usually distortion-free samples that are quite different from actual content for media services. Video content is normally compressed with lossy video encoders so quantization loss inevitably arises. In addition, care for the complexity of deep neural networks should be taken for real media services. Therefore, we pursue this point so that a scheme using actual data from media content providers when training the network is proposed in this paper.

i) Generating dataset

In order to efficiently convert SD videos to FHD videos, we first prepare a training dataset pair (input-target) and then train the proposed network. Figure 3 shows the overall process of the proposed up-scaling scheme.

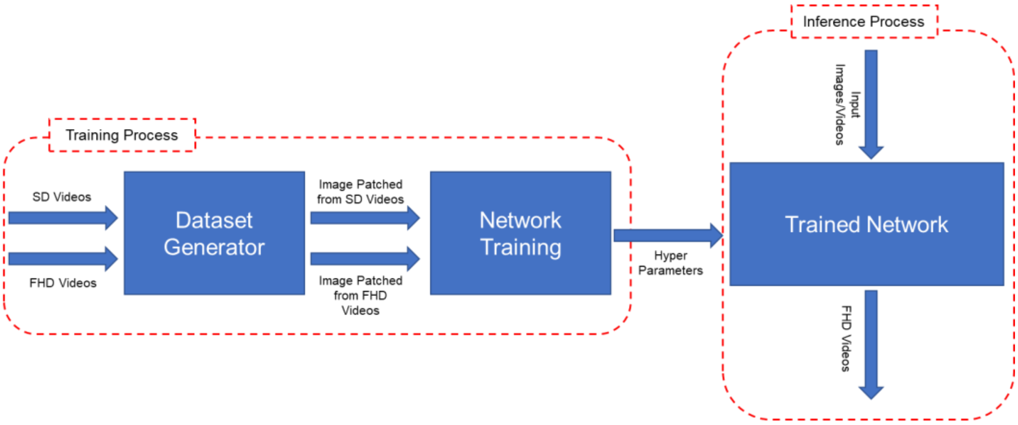
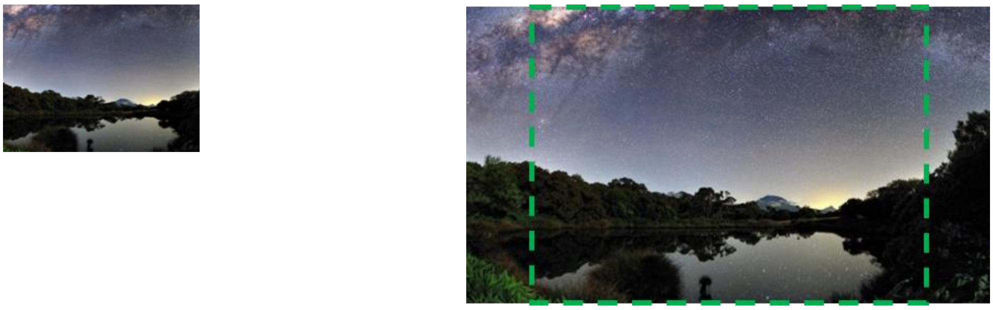


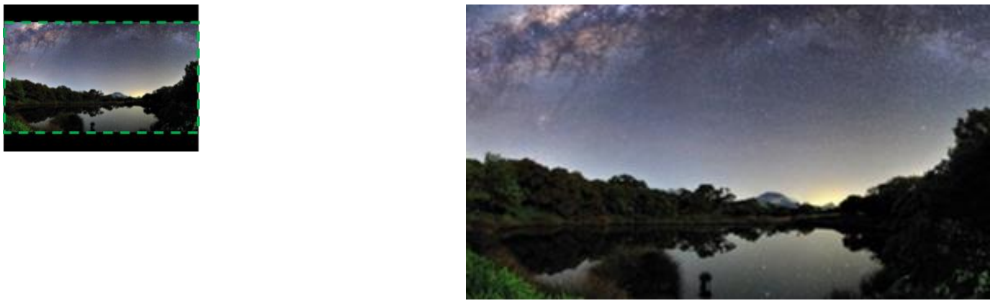
Figure 3. The overall process of the proposed up-scaling scheme

In Figure 3, the Dataset Generator automatically generates training data pairs with SD (720x480) videos and FHD (1920x1080) videos, respectively. Unfortunately, two image patches extracted from the same SD videos and FHD videos usually do not exactly match due to different aspect ratios. Since, the ratios of SD videos and FHD videos are 4:3 and 16:9, respectively, it is essential that images to construct the training pair from SD videos are aligned with those from FHD videos before training the network. Figure 4 shows an example of how this alignment process begins in the Dataset Generator.



(a) SD image without black padding area

(b) Cropped FHD image



(c) Cropped SD image with black padding

(d) Original FHD image

Figure 4. An example how the alignment process between SD and FHD begins in the Dataset Generator

It is mostly observed that commercial SD videos contain either a black padding area or not. When SD videos have no black padding areas, FHD videos are cut to fit the 4:3 ratio as shown in Figure 4-(b). In case of SD videos with black padded areas, the video to fit the 16:9 ratio should be cropped. Then, images from SD videos are upscaled by 2 using a bi-cubic interpolation method. Finally, upscaled SD images are aligned with the cropped FHD images using the previous alignment method [10].

ii) Proposed deep neural network for up-scaling

Figure 5 shows our proposed network model that has 8 long residual blocks (LRBs) for up-scaling process.

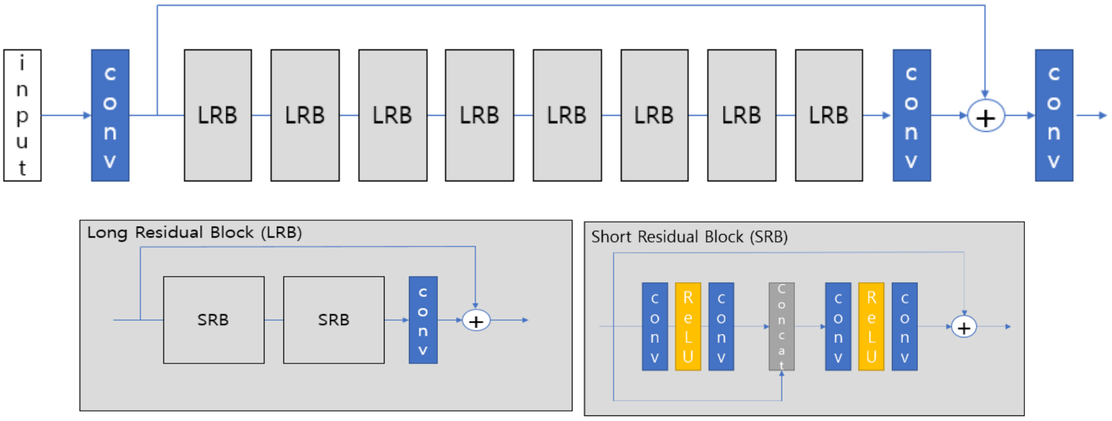


Figure 5. Proposed network model that has 8 LRBs for up-scaling process

As shown in Figure 5, the LRBs consists of two short residual blocks (SRBs), a skip connection and one additional convolution layer. Each SRB concatenates the output feature map of the two convolution layers and the input feature map of the SRB and transfers it to the next convolution layer. In addition, the skip connection of the SRB serves to transfer the features of the input feature map to the next SRB by adding the input feature map to the output feature map. In order to focus on the restoration of the high-frequency region of the image, the concat block as shown in Figure 5 is inserted, and four convolutions were repeated without repeating two convolutions. By using LRB as well as SRB, skip connections are connected to each of the 9 layers so that learning can be done even when the number of layers becomes deeper.

2. HFR module

High Frame rate (HFR) can improve the visual quality by generating intermediate video frames between two existing consecutive frames. In general, HFR is a very challenging problem when accurately interpolating fast motion frames. We introduce a convolutional

Long-Short Term Memory (LSTM) and Convolutional Neural Network (CNN) based HFR method to accurately interpolate fast motion frames by effectively capturing the temporal dynamics of fast local and global motion. Our model can handle temporal information while processing the spatial information of images with 3×3 convolution filters. Our HFR method learns to generate an intermediate frame between two consecutive input frames, the previous frame and the next frame. Figure 6 shows the architecture of our proposed HFR method.

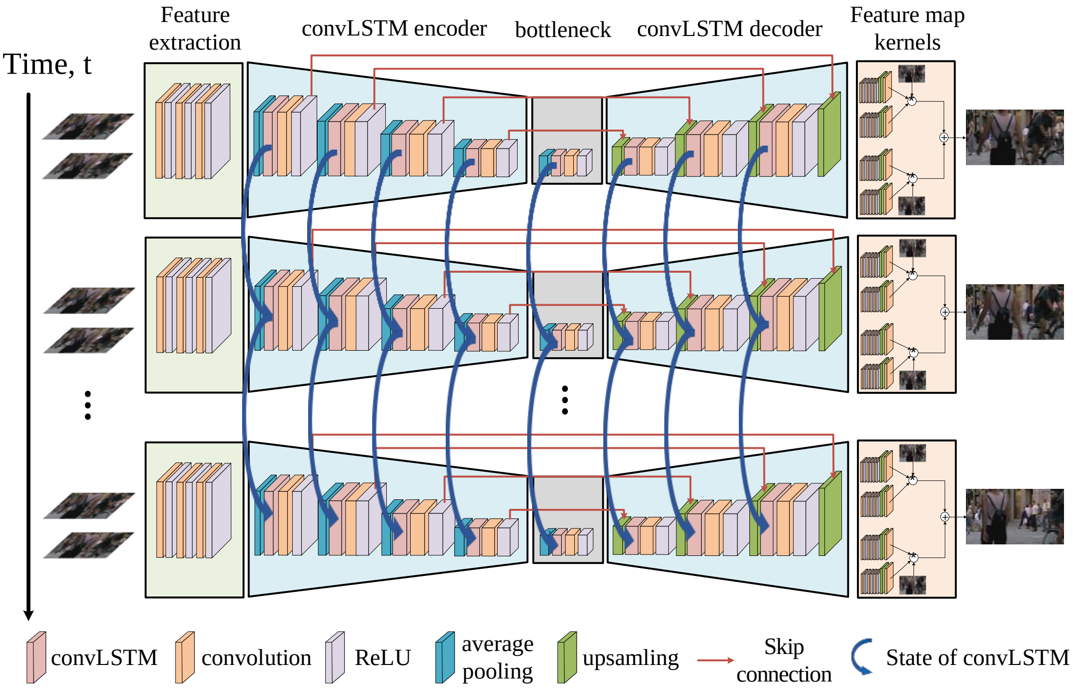


Figure 6. The architecture of our proposed HFR

This model consists of the following five modules: the front-end feature extraction module, the convolutional LSTM (convLSTM) encoder module, the bottleneck feature map module, the convLSTM decoder module and the feature-map-driven kernel module, which are cascaded in sequential order. The CNN-based feature extraction module takes as input the two consecutive frames the previous frame and the next frame, and generates the features maps as input for the convLSTM encoder that can simultaneously process spatial and temporal information between two consecutive frames to capture the spatio-temporal dynamics together with the convLSTM decoder in compact feature domains. The CNN-based feature extraction network and the convLSTM uses 3×3 convolution filters. The CNN-based feature-map- driven kernel estimation network focuses to construct feature maps that are used to estimate both the horizontal and vertical kernels for the previous frame and the next frame. The CNN- based feature-map-driven kernel estimation network produces the final intermediate frame by convolving the previous and next input frames with the generated four feature maps (two separable horizontal and vertical kernels for the previous frame, and two separable horizontal and vertical kernels for the next frame) of its last convolution layer. Figure 7 describes the feature-map-driven kernel operation.

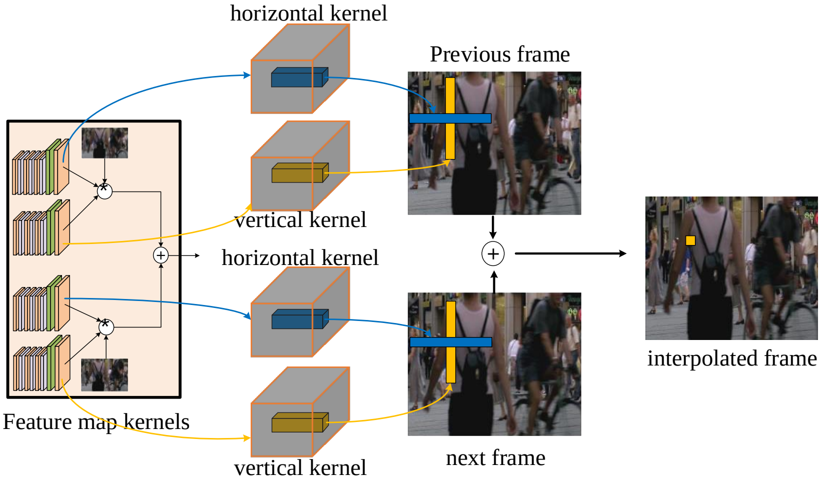


Figure 7. The feature-map-driven kernel operation

For the previous frame, a 1-D horizontal kernel performs a convolution operation in the horizontal direction for each pixel and then a 1-D vertical kernel performs a convolution on the output of horizontal convolution operation. The same process is performed for the next frame and then generates an interpolated frame.

3. Re-targeting module

We also propose a deep learning-based image/video retargeting method that converts original image/video whose aspect ratio is fixed to re-scaled images or videos of a desired aspect ratio. This results in maximizing the utilization of the display and minimizing the perception of distortion compared to applying linear scaling methods on the original image/video. Our approach is to utilize the re-targeting network with user's content consumption circumstance as Control Parameters as depicted in Figure 2, such as display size and display viewing mode, i.e. landscape or portrait. In other words, the re-targeting module in SUPERNOVA provides users with a re-scaled image/video without perceptional loss regardless of aspect ratio and viewing mode of the various displays.

The proposed re-targeting method consists of a saliency detection part and a re-sizing operation part that considers the aspect ratio of the display and the users' viewing mode. The Saliency detection part is depicted in Figure 8.

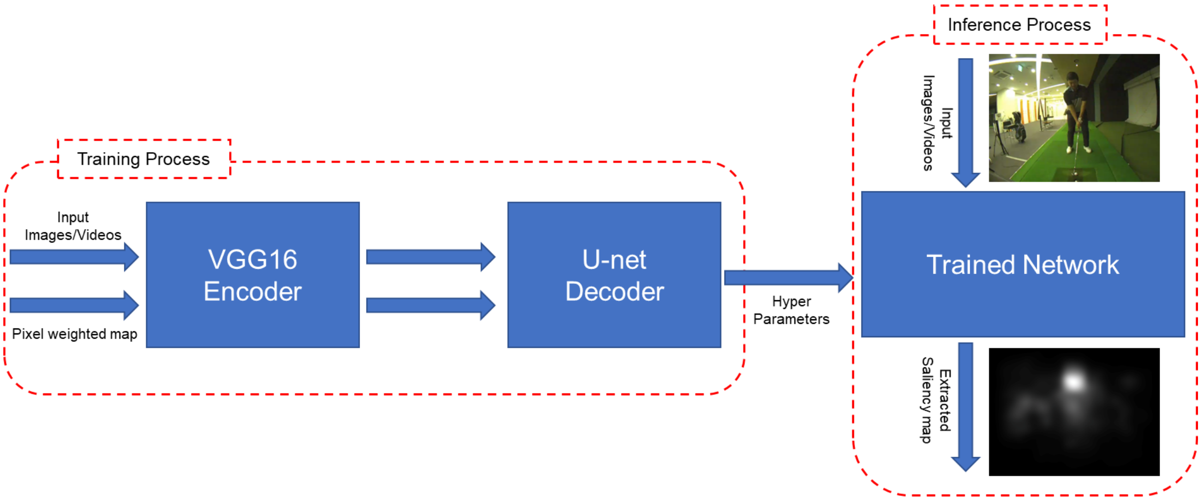


Figure 8. The feature-map-driven kernel operation

To decide the saliency region from the original image/video, we first combine VGG16 [11] encoder and U-net [12] decoder and train this combined network with the Semantic-Retarget dataset [13], where the pixel relevance score was annotated by several subjects. Due to the small size of the Semantic-Retarget dataset, we duplicated the amount of data using an augmentation process. For the re-sizing operation part, we applied results of the extracted saliency map to the previous study [7].

Section IV: FEASIBILITIES FOR NON-MEDIA APPLICATIONS WITH SUPERNOVA

A. Nano Wafer Pattern Image De-noising

A de-noising module for nano wafer pattern images was recently added. This deep learning-based image enhancement solution, can be applied to the semiconductor inspection process. High-quality semiconductor image acquisition is essential for QA processes that determine defects in semiconductor wafers. The inspection equipment used in the semiconductor QA process basically emits high voltage electrons to the wafer to image the reflectance. However, due to the non-uniformity of reflectance, a low-quality image containing severe noise is obtained in a single shot. To solve this, the existing semiconductor inspection equipment undergoes a step of repeatedly photographing a physically identical position of the wafer several times and synthesizing an image to remove noise. High-quality images can be obtained through the above process, but it takes time for the image synthesis, reducing the Turn Around Time (TAT) throughput of the inspection process. SUPERNOVA dramatically reduces the number of shots and the high-quality image acquisition time by removing noise from a single shot image at high speed through a deep learning algorithm. Figure 9 shows that the proposed deep neural network for noise reduction.

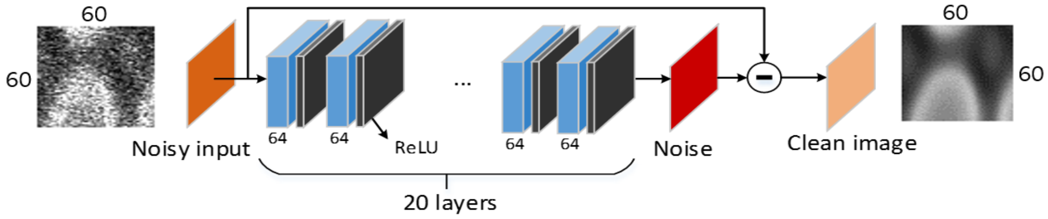


Figure 9. The proposed deep neural network for noise reduction

As depicted in Figure 9, The proposed network uses a residual learning-based network to remove the noise in the input image, and generates noise included in the input image through the network, and then removes it from the input image to generate an image with the target noise removed. A total of 20 convolutional layers were used, and the output of 19 layers excluding the last layer passes ReLU activation. The convolution filter size is 3x3, and the number of filters in each layer is 64. The loss function used to train the network was taught to minimize the square of the difference between the input image and the output image with L2 loss.

B. Face image restoration

Another deep learning-based method for restoring face images was also added in SUPERNOVA, and has significantly improved visual quality in various fields from CCTV image to broadcasting content. Moreover, it is crucial to restore the facial image to its best state when remastering an old video because viewers focus on a person, especially the face, rather than the surrounding background. In this paper, a GAN-based face image restoration method was proposed for improving the facial region in low-resolution and low-quality image including quantization error, camera noise and blurring.

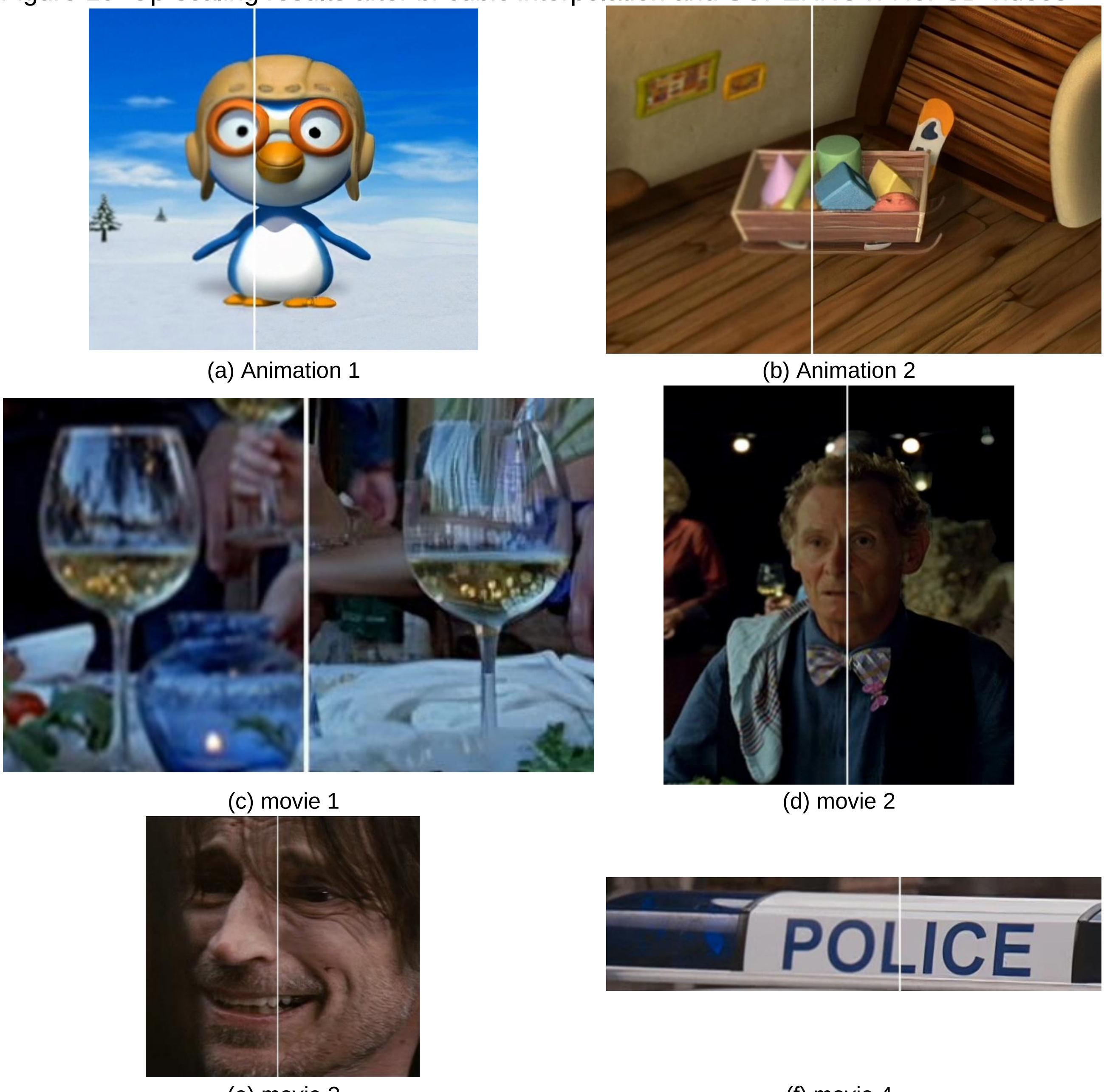
In order to apply GAN-based image restoration, it is essential to consider the stability for the training process and the variety of the training dataset. In this paper, GAN [15], composed of the generative network and the discriminative network, is incorporated with the gradient penalty and spectral normalization to achieve stable training. FaceNet [8] also specialized in face recognition and analysis is also used to reflect the characteristics of the face in the objective function.

For robust output in the deep learning process, the variety of the training dataset should be prepared. In this paper, FFHQ [9] dataset was used to train. This dataset consisted of 70,000 face images at 1024x1024 with a variation in age, ethnicity and image background and accessories such as sunglasses, hats, eyeglasses, etc. In the training process, the face is aligned by using a facial landmark to improve performance in the GAN. In addition, the training is conducted considering white noise, blurring effect, and quantization error to ensure a robust performance.

Section V: PERFORMANCE

To verify the effectiveness of our proposed SUPERNOVA with its 3 modules, we present some images and PSNRs before and after SUPERNOVA. Each network configuration is already mentioned in Section III and details of the experimental conditions are omitted in this section. Figure 10 shows up-scaling results before and after SUPERNOVA for SD videos.

Figure 10. Up-scaling results after bi-cubic interpolation and SUPERNOVA for SD videos.



(a) Animation 1

(b) Animation 2

(c) movie 1

(d) movie 2

(e) movie 3

(f) movie 4

As can be seen from Figure 10, up-scaled SD images by bi-cubic interpolation are found in the left-side and up-scaled SD images by SUPERNOVA are found in the right-side. It is clear that images after SUPERNOVA show better performance than those after bi-cubic interpolation. Since, we only treat original SD contents where original FHD contents do not exist, reference-based quality assessment metrics cannot be used for comparison.

Another experimental result for the proposed HFR method is presented in Table 2.

Table 2. An experimental result for the proposed HFR method with PSNR

Video Sequences	Method in [15] (dB)	Method in [16] (dB)	Method in [5] (dB)	SUPERNOVA – HFR module (dB)
Pedestrian Area	30.11	28.22	30.94	**31.46**
Tractor	29.09	26.49	29.67	**30.18**
See You Again	38.84	39.04	40.91	**41.31**
Average	32.68	31.25	33.84	**34.22**

Five FHD (1920x1080) video sequences were acquired from the Tom Scott channel on YouTube to construct a training data set. After we resized them to 1280x720 resolution, we randomly took training samples from the resized five video sequences where each training sample consists of 10 consecutive frame patches of 256x256x3 size having 3 RGB channels. For the experiments, the training samples that contained motion larger than 15 pixels were selected based on estimated optical flow values. The selected training samples were randomly flipped horizontally and vertically for robust training. Note that when a training sample is flipped, all 10 consecutive patches are simultaneously rotated by 180 degree. We use zero-adding before convolutions to keep the sizes of the feature maps the same. It is interesting to note that the performance does not fluctuate whether using clean dataset or actual dataset.

Finally, some results after re-targeting module is presented in Table 3.

Figure 11. Comparisons of source images (a,d), linear scaled images (b,e) and our re-targeted images (c,f) in landscape mode (a, b, c) and portrait mode (d, e, f).

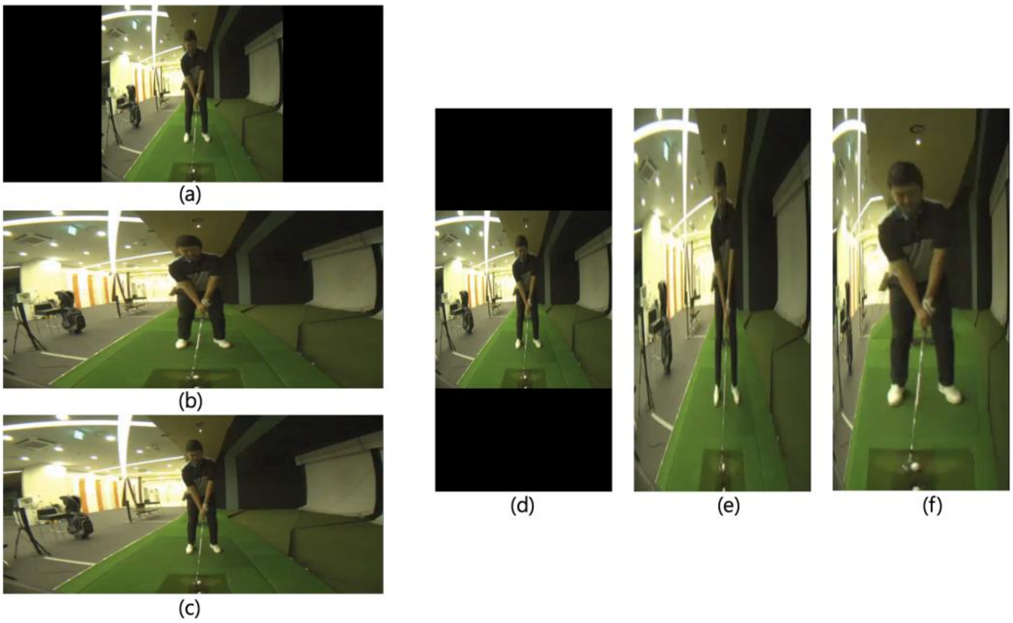
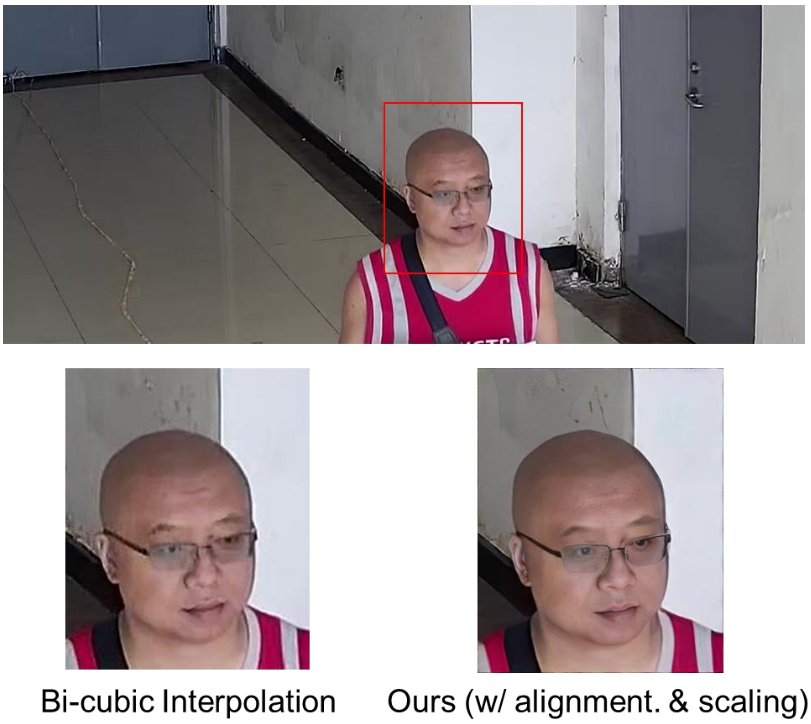


Figure 11 shows the original image with 4:3 aspect-ratio is converted into images with 19:9 (landscape mode) and 9:19 (portrait mode), respectively. It obviously shows that the image ratio for the salient region is well maintained when applying our re-targeted method, while the distortion of the salient region is perceivable in the linear scaled image.

Section VI: CONCLUSIONS and Future Work

In this paper, a media quality enhancement platform with its component methods were proposed. There are 3 methods in the platform, which are for up-scaling, HFR and re-targeting, respectively. For up-scaling, we first introduced a pre-processing to efficiently prepare the training dataset and then proposed a novel deep neural network for better performance. For HFR, we proposed a novel structure for deep neural network and showed its verification with PSNR results compared to other previous methods. For re-targeting, we extracted saliency gray-scale map with the proposed scheme. This map was useful to generate image pixels for black areas of original image/video. From all these methods, it is apparent that the image/video quality significantly increases after SUPERNOVA. Our future work is to implement more functions in the current SUPERNOVA platform. De-noising and face image restoration are being tested, and colorization and de-fogging methods are to be considered. Last, it is noticeable to see Figure 12 because the texture and shape of the original image is well restored, and we hope these technologies can restore extremely bad quality of CCTV images to settle unsolved criminal cases.


Bi-cubic Interpolation

Ours (w/ alignment. & scaling)

Figure 12. Performance comparison after bi-cubic interpolation and SUPERNOVA

REFERENCES

1. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

2. Ledig, Christian, et al. "Photo-realistic single image super-resolution using a generative adversarial network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

3. Lim, Bee, et al. "Enhanced deep residual networks for single image super-resolution." Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2017.

4. Zhang, Yulun, et al. "Image super-resolution using very deep residual channel attention networks." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

5. Niklaus, Simon, Long Mai, and Feng Liu. "Video frame interpolation via adaptive separable convolution." Proceedings of the IEEE International Conference on Computer Vision. 2017.

6. Liu, Ziwei, et al. "Video frame synthesis using deep voxel flow." Proceedings of the IEEE International Conference on Computer Vision. 2017.

7. Rubinstein, Michael, Ariel Shamir, and Shai Avidan. "Multi-operator media retargeting." ACM Transactions on graphics (TOG) 28.3 (2009): 1-11.

8. Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

9. Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

10. Szeliski, Richard. "Image alignment and stitching: A tutorial." Foundations and Trends® in Computer Graphics and Vision 2.1 (2007): 1-104.

11. Han, Song, Huizi Mao, and William J. Dally. "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding." arXiv preprint arXiv:1510.00149 (2015).

12. Jansson, Andreas, et al. "Singing voice separation with deep U-Net convolutional networks." (2017).

13. Liu, Si, et al. "Composing semantic collage for image retargeting." IEEE Transactions on Image Processing 27.10 (2018): 5032-5043.

14. Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems. 2014.

15. Tsai, Tsung-Han, An-Ting Shi, and Ko-Ting Huang. "Accurate frame rate up-conversion for advanced visual quality." IEEE transactions on broadcasting 62.2 (2016): 426-435.

16. Meyer, Simone, et al. "Phase-based frame interpolation for video." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

ACKNOWLEDGEMENTS

As the main authors, Taeyoung Na and Jin Jeon equally contributed to this paper.