Information Fusion based Quality Enhancement for 3D Stereo Images Using CNN

Zhi Jin[1 3*] , Haili Luo[1*], Lei Luo[2], Wenbin Zou[1] , Xia Li[1], Eckehard Steinbach[3]
1 College of Information Engineering, Shenzhen University, Shenzhen, P.R.China.
2 College of Telecommunication and Information Engineering,
Chongqing University of Posts and Telecommunications, Chongqing, P.R.China.
3 Technical University of Munich, Chair of Media Technology, Munich, Germany

***Abstract*—Stereo images provide users with a vivid 3D watching experience. Supported by per-view depth maps, 3D stereo images can be used to generate any required intermediate view between the given left and right stereo views. However, 3D stereo images lead to higher transmission and storage cost compared to single view images. Based on the binocular suppression theory, mixed-quality stereo images can alleviate this problem by employing different compression ratios on the two views. This causes noticeable visual artifacts when a high compression ratio is adopted and limits free-viewpoint applications. Hence, the low quality image at the receiver side needs to be enhanced to match the high quality one. To address this problem, in this paper we propose an end-to-end fully Convolutional Neural Network (CNN) for enhancing the low quality images in quality asymmetric stereo images by exploiting inter-view correlation. The proposed network achieves an image quality boost of up to 4.6dB and 3.88dB PSNR gain over ordinary JPEG for QF10 and 20, respectively, and an improvement of up to 2.37dB and 2.05dB over the state-of-the-art CNN-based results for QF10 and 20, respectively.**

I. Introduction

Stereo images have witnessed a revived interest in recent years. Associated with per-view depth maps, they support free navigation into other viewpoints by view synthesis techniques, such as Depth-Image-Based Rendering (DIBR) [1]. However, compared with single view images, 3D stereo images have stronger requirements on the acquisition, storage and transmission units of multimedia systems. Therefore, efficient image compression schemes are highly demanded. Referring to the binocular suppression theory [2] of the human vision system, which states that the perceived stereo quality is mainly determined by that of the high quality view in the binocular vision system, asymmetric compression can be adopted for stereo images, where two views are encoded with different spatial resolutions (resolution-asymmetry) or different PSNR qualities (quality-asymmetry) [3]. Having the same resolution, quality asymmetric stereo images are straight-forward to use to generate intermediate views, which leads to the widely adopted quality-asymmetric compression format on stereo images.

There are two common categories for image compression: lossless (e.g., PNG) and lossy methods (e.g., JPEG [4]). Although lossy compression leads to a non-invertible information loss, it can achieve a much higher compression ratio. However, users have to pay for the undesired artifacts, which severely reduce the watching experience. Taking JPEG as an example, by adopting Block-based Discrete Cosine Transform (BDCT) together with coarse quantization, JPEG compression aims at reducing the inter-pixel statistical redundancy to achieve high compression ratio. However, this gives rise to the discontinuity of intensities between adjacent blocks, which is called blocking artifacts, and the truncation of High Frequency (HF) BDCT coefficients, which results in ringing artifacts.

Although quality asymmetric compression can reduce the transmission cost, the adoption of a high compression ratio limits the suitability for free-viewpoint applications. Based on *Saygili* and *Aflaki's* subjective quality assessments [5] [6], there is a "just-noticeable asymmetric PSNR threshold" for stereo images. If the PSNR of the Low Quality (LQ) view is below this threshold, the degradation of stereo image quality can be perceived even if the High Quality (HQ) view maintains high PSNR quality. For different kinds of displays, the threshold values are slightly different (31dB for parallax barrier displays and 33dB for full resolution projection displays [5]). This dilemma, however, can be reduced by enhancing the quality of the low quality view to match the high quality one at the receiver side. To achieve this, different approaches have been proposed. Pointwise Shape-Adaptive DCT (SA-DCT) [7] as a deblocking method relies on attenuated DCT coefficients to reconstruct a local estimate of the signal within the adaptive-shape support. However, the improved images still lack of sharp edges, and have overly smoothed texture regions. The Regression Tree Fields (RTF) based restoration method [8] starts from the SA-DCT results and produces globally consistent image reconstructions with a regression tree field model. However, both SA-DCT and RTF approaches remove the blocking artifacts while blurring the images. During the last ten years, deep learning based approaches have led to a series of breakthroughs in computer vision and have been proven to be also a powerful tool for low-level vision problems. By learning a nonlinear mapping between compressed and original images, they obtain the state-of-the-art results. *Dong et al.* [9] introduced a 4-layer CNN (ARCNN) to learn an end-to-end mapping between the LQ and the original images, which achieves a significant quality improvement for the LQ images. Extended from [9], *Yu et al.* [10] proposed a faster 5-layer CNN, called FastARCNN.

In quality-asymmetric stereo images, due to the inter-view

E-mail: jinzhi_126@163.com; wzouszu@sina.com.

* means equal contribution. Corresponding author: Wenbin Zou.

ISBN 978-90-827970-1-5 © EURASIP 2018

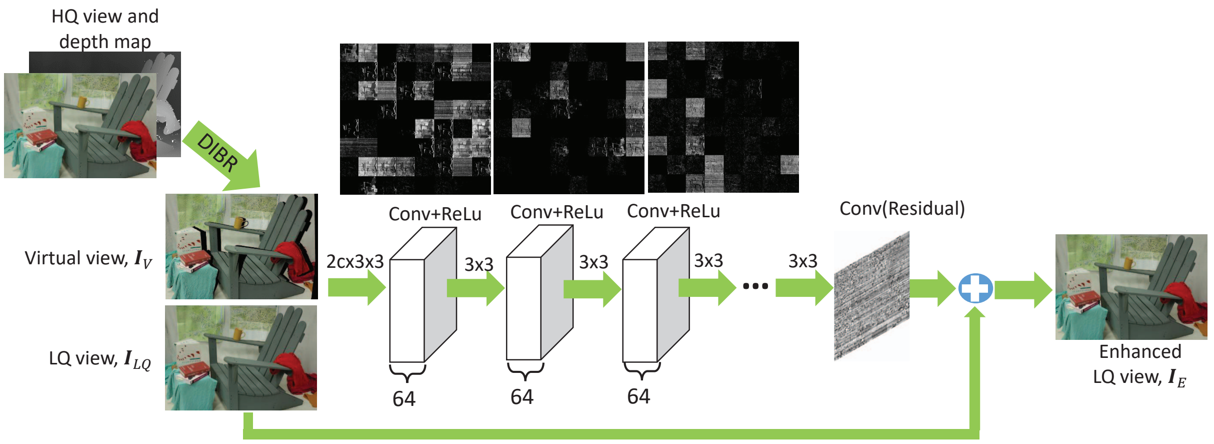

HQ view and depth map
DIBR
Virtual view, $\boldsymbol{I}_V$
LQ view, $\boldsymbol{I}_{LQ}$
Conv+ReLu
Conv+ReLu
Conv+ReLu
Conv(Residual)
2cx3x3
3x3
3x3
3x3
3x3
64
64
64
Enhanced LQ view, $\boldsymbol{I}_E$

Figure 1. The architecture of the proposed fully convolutional neural network.

redundancy, the retained HF information in the HQ images is capable to recover the lost HF information in the LQ images. Hence, we propose to use an end-to-end fully convolutional neural network to extract the corresponding HF information from the HQ image and fuse it with the LQ image. To avoid binocular disparity caused negative effects, the virtual view image generated from the HQ image and the depth map at the viewpoint of the LQ image is combined with the LQ image as the two inputs of the proposed network. The information fusion happens at the first convolutional layer. Residual learning is adopted to boost up the training process and reconstruction accuracy. Extensive experiments indicate our approach can efficiently enhance the quality of the LQ images as compared to the state-of-the-art approaches.

II. Proposed Quality Enhancement Network

During compression, by using a smaller compression ratio on the HQ view, more HF information is preserved. With the observation that the two stereo views contain a large amount of redundant information, the HQ view and its corresponding depth map can be exploited to generate the virtual view at the viewpoint of the LQ view and utilized to enhance the quality of the LQ view. Assuming that the warping process works accurately and maps the HQ pixels into the virtual view without introducing tangible warping distortion, the virtual view can inherit most of the HF information from the HQ view. However, due to the change of viewpoint, the generated virtual view includes some missing parts, and usually these parts are called “occluded regions” or “holes”. In our approach, without any post-processing, the virtual view image and the LQ image are treated as the two inputs of the proposed network, and all the useful information from the HQ view (or virtual view) will be fused with the LQ view by the proposed CNN.

A. Network Architecture

Fig. 1 shows the architecture of the proposed CNN [1]. The proposed CNN consists of n convolutional layers, the size and number of convolutional filters for each layer are set to 3×3 and 64, respectively.

The LQ image $\mathbf{I}_{LQ}$ and the virtual view image $\mathbf{I}_V$ are concatenated as the inputs of the first layer, so that there are $2c$ input channels, where c is the number of input channels for each image. The first layer is used for **feature extraction and information fusion**, and generates 64 feature maps. By adding a bias term to the outputs and employing ReLU as the activation function, the convolution operation in this layer can be formulated as:

$$F_1(\mathbf{I}_{LQ}, \mathbf{I}_V) = ReLU(W_{1,1} * \mathbf{I}_{LQ} + W_{1,2} * \mathbf{I}_V + B_1) \quad (1)$$

where W and B represent the weights and biases, respectively; ‘$*$’ denotes the convolution operation; F represents the nonlinear mapping process.

It is worth noting that in this step, the features of the occluded regions in $\mathbf{I}_V$ are also extracted and fused into the output feature maps, which introduce some negative effects on the reconstructed image. However, these effects are hugely suppressed by the following convolutional layers (Fig. 4), which can be regarded as enhancing the extracted features. Different from traditional methods, such as [12], where linear mapping was adopted to enhance the quality of the LQ images, in the proposed network, the intermediate $n-2$ layers are used to nonlinearly map generated feature maps. Hence, these layers are used for **feature enhancement and non-linear mapping**. The last layer is used for **image reconstruction**. After the first $n-1$ convolutional layers, features in both inputs are successfully extracted, fused and enhanced by optimizing the output of the network to generate the final enhanced LQ image. The operation of these layers can be formulated as:

$$F_i(\mathbf{I}) = ReLU(W_i * F_{i-1}(\mathbf{I}) + B_i) \quad (2)$$

where i indicates the ith convolutional layer; $F_i(\mathbf{I})$ and $F_{i-1}(\mathbf{I})$ are the outputs of the ith and $i-1$th convolutional layer, respectively; W_i and B_i represent the weights and biases for the ith layer, respectively.

Since residual learning is adopted in the proposed work, the final learned residual map containing the restored HF details is expressed as

$$F_n(\mathbf{I}) = W_n * F_{n-1}(\mathbf{I}) + B_n \quad (3)$$

[1] The target of the proposed CNN is to enhance the texture quality of the LQ view, so that we do not show the corresponding compressed depth map of the LQ view in this figure. However, for the quality enhancement of the compressed depth maps, please refer to our work [11].

Table I
QUALITY ENHANCEMENT COMPARISON WITH THE STATE-OF-THE-ART ALGORITHMS ON MIDDLEBURY AND MPEG DATA

Dataset	Algorithms	q3		q5		q7		q10		q20		#Para	Runtime
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	(k)	(second)
Middlebury data	JPEG [4]	27.19	0.8322	30.28	0.8582	32.51	0.8845	34.68	0.9065	38.28	0.942	–	–
	SA-DCT [7]	28.41	0.8801	31.83	0.9092	34.20	0.9270	36.52	0.9768	39.94	0.993	–	–
	ARCNN [9]	29.27	0.8911	32.57	0.9166	34.76	0.9300	36.69	0.9422	39.85	0.959	106	0.266
	FastARCNN [10]	–	–	–	–	–	–	36.91	0.9426	40.11	0.960	56	0.146
	Late-4	33.45	0.9452	35.38	0.9534	36.80	0.9579	38.36	0.9635	41.44	0.973	112	0.370
	Proposed-4	33.52	0.9459	35.60	0.9551	37.02	0.9602	38.60	0.9653	41.29	0.972	75	0.239
	Proposed-8	**33.79**	**0.9490**	**36.15**	**0.9588**	**37.72**	**0.9637**	**39.28**	**0.9681**	**42.16**	**0.975**	223	0.479
MPEG video data	JPEG [4]	27.34	0.8112	30.06	0.8424	31.91	0.8645	33.72	0.8886	36.92	0.9262	–	–
	SA-DCT [7]	28.57	0.8573	31.48	0.8880	33.43	0.9051	35.27	0.9632	38.28	0.983	–	–
	ARCNN [9]	29.19	0.8646	31.99	0.8932	33.80	0.9086	35.53	0.9238	38.30	0.9428	106	0.075
	FastARCNN [10]	–	–	–	–	–	–	35.68	0.9240	38.51	0.9436	56	0.033
	Late-4	34.34	**0.9404**	35.43	0.9451	36.30	0.9476	37.51	0.9532	39.77	0.9610	112	0.067
	Proposed-4	34.32	0.9391	35.67	0.9453	36.72	0.9495	37.72	0.9536	39.70	0.9608	75	0.060
	Proposed-8	**34.46**	0.9399	**36.03**	**0.9463**	**37.08**	**0.9505**	**38.14**	**0.9544**	**40.20**	**0.9623**	223	0.090

Moreover, the final enhanced LQ image can be expressed as:

$$F(\mathbf{I}_{LQ}, \mathbf{I}_V) = F_n(\mathbf{I}) + \mathbf{I}_{LQ} \tag{4}$$

Aiming to learn the optimal values for the weight and bias of each layer, the L2 distance between the reconstructed images $F(\mathbf{I}; \Theta)$ and corresponding uncompressed ground truth images $\mathbf{I}_{GT}$ needs to be minimized. Given a set of m uncompressed images $\mathbf{I}^i_{GT}$, the corresponding compressed images $\mathbf{I}^i_{LQ}$ and generated virtual view images $\mathbf{I}^i_V$, the Mean Squared Error (MSE) loss function is:

$$Loss(\Theta) = \frac{1}{m} \sum_{i=1}^{m} \| F(\mathbf{I}^i_{LQ}, \mathbf{I}^i_V; \Theta) - \mathbf{I}^i_{GT} \|_2^2 \tag{5}$$

where Θ contains the parameters of the network, including both weights and biases.

B. Information Fusion Order

In the proposed network, since the information fusion happens at the first convolutional layer, it is called "Early fusion". An alternative network structure is shown in Fig. 2, which can be called "Late fusion". In the late fusion network, the two input images are fed separately into a convolutional layer. Then the corresponding feature maps are concatenated and fed into the later layers. Therefore, the information fusion happens later. Compared with the early fusion, since more convolutional layers are involved, the late fusion structure has more parameters that need to be trained. Moreover, suffering from more negative effects from the occluded regions, the late fusion network has worse performance than the proposed early fusion network, however, it is still superior to the state-of-the-art networks. The details are discussed in Sec.III-C.

III. EXPERIMENTS

In this section, we first present the implementation details and then compare the proposed network with the state-of-the-art networks on two multiview datasets quantitatively and qualitatively. Finally, we discuss the effects of fusion order in the proposed network.

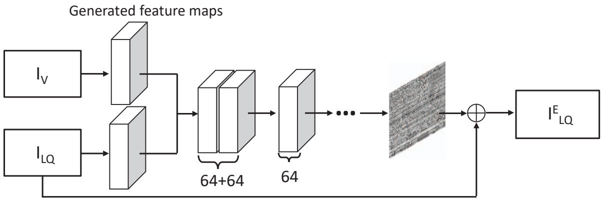


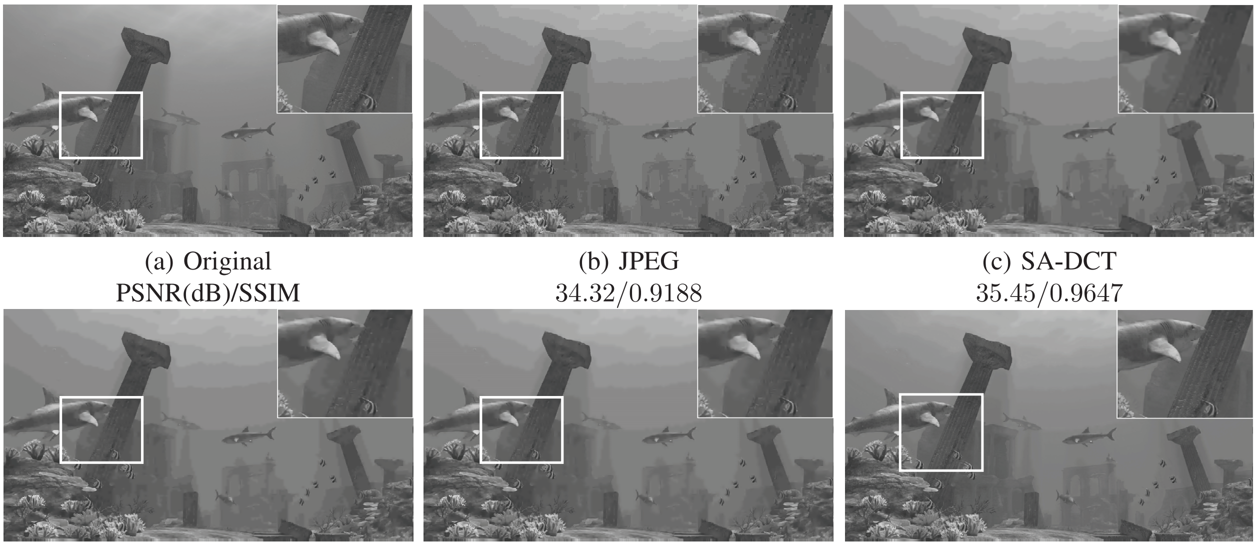
Figure 2. The architecture of the late information fusion network.

A. Implementation Details

To train the proposed network, we adopt the Middlebury 2014 stereo image dataset [13] containing 18 images of size 2964×1988 as our training data. The remains 5 stereo images and 7 first frames of the MPEG multivew video sequences are used to validate the network's performance. Taking into account the training complexity, small patch training strategy is adopted, so that training images are split into 24×24 patches with the stride of 10 and there are $1,008,128$ training sub-images in total. In this work, Adam [14] solver is used with the batch size 128, the first and second momentum parameters are set to 0.9 and 0.999, respectively, and the weight decay is 10^{-4}. The initial learning rate is 0.001 and decays every 10 epochs by a factor of 10. There are 50 epochs in total. In this work, we propose two similar network architectures which differ only in their network depth. We distinguish $n = 4$, referred to "Proposed-4" and $n = 8$, referred to "Proposed-8" in this section. Moreover, we focus on the quality enhancement of the image luminance, and hence, in the proposed framework $c = 1$.

B. Comparison with the State-of-the-Art Approaches

In order to evaluate the performance of the proposed network, especially when the quality of the LQ view is below the just-noticeable asymmetric PSNR threshold, JPEG quality $QF = 3, 5, 7$ are used to generate the highly compressed LQ images whose PSNR quality is in the range of 27dB to 32dB, and the virtual views are generated from the HQ view and its corresponding depth map without any compression. The



(a) Original
PSNR(dB)/SSIM

(b) JPEG
34.32/0.9188

(c) SA-DCT
35.45/0.9647

(d) ARCNN
35.86/0.9446

(e) FastARCNN
35.97/0.9440

(f) Proposed 8-layer
39.93/0.9806

Figure 3. Visual comparison results that achieved with the proposed CNN and other state-of-the-art approaches on MPEG image "Shark".

corresponding experimental results on the two datasets are shown in Table I. The results in bold indicate the highest values among all the results. Furthermore, the number of network parameters for the deep learning based approaches are also given. From these results, it is clear that for all cases, the LQ image quality is improved above the highest display threshold by both of the proposed networks, which means after enhancement the quality degradation of stereo images are not perceived. Consequently, the usable range for the quality-asymmetric compression method has been extended. For the "below threshold" LQ images at QF3, SA-DCT also can improve their quality, however, it fails in improving the quality above its nearest display threshold, which is similar as ARCNN. By retraining the ARCNN model with the provided hyper-parameters, we can perceive that the proposed network outperforms ARCNN for the three tested QFs. Although our 4-layer CNN can improve the image quality a lot, compared with the proposed 8-layer one, its results are still slightly lower. However, the 4-layer structure has less network parameters than the 8-layer one. Hence, there is a trade-off between performance and computational cost, which is confirmed by the runtime comparison reported in Table I.

The comparison results of the average image quality enhancement at JPEG quality 10 and 20 are also shown in Table I. For all the datasets and JPEG qualities, our network gives the highest PSNR and SSIM results over other state-of-the-art approaches, and the PSNR gains are up to 2.37dB and 2.05dB over the best previous learning based results obtained by FastARCNN on Middleburry dataset for QF10 and 20, respectively. The PSNR gains are up to 2.46dB and 1.69dB over FastARCNN on MPEG dataset for QF10 and 20, respectively. With the same network depth, the proposed 4-layer network has 30% less parameters but still outperforms ARCNN. Moreover, Fig. 3 shows one visual comparison result, and we can observe that the HF information in the virtual view image has been successfully extracted and fused with the LQ image.

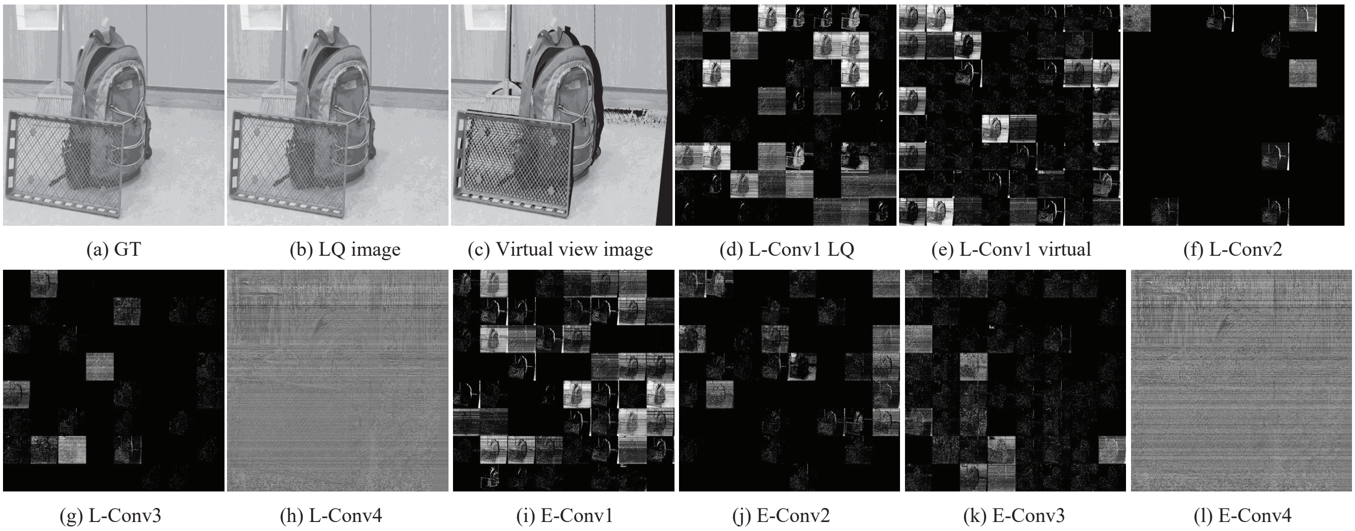
In this subsection, besides the two widely used full-reference 2D Quality Assessment (QA) metrics, i.e., PSNR, and SSIM, we also evaluate the performance of the proposed network with two binocular frequency integration 3D QA metrics, i.e., Frequency Integrated PSNR (FI-PSNR) and Frequency Integrated SSIM (FI-SSIM) [15]. The FI-metrics incorporate binocular behavior into 2D objective metrics for evaluating the quality of stereo images, and the FI-metrics can achieve performance consistency with the corresponding subjective Mean Opinion Scores (MOS) [15]. Corresponding results are shown in Table II. In Table II, the performance of the proposed 8-layer model is highest among all the other three stereo models in the 3D average metrics. The proposed 4-layer model has the second best performance.

Table II
THE BINOCULAR PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART APPROACHES

QF		Algorithms	FI-PSNR	FI-SSIM		Algorithms	FI-PSNR	FI-SSIM
10	Middlebury data	JPEG [4]	45.72	0.7986	MPEG video data	JPEG [4]	45.77	0.8231
		SA-DCT [7]	45.94	0.8058		SA-DCT [7]	45.98	0.8315
		ARCNN [9]	46.49	0.8147		ARCNN [9]	46.47	0.8491
		Proposed-4	47.75	0.9109		Proposed-4	48.29	0.9337
		Proposed-8	**48.57**	**0.9144**		Proposed-8	**49.66**	**0.9352**
20		JPEG	52.65	0.8742		JPEG	52.54	0.8869
		SA-DCT [7]	52.91	0.8782		SA-DCT [7]	52.82	0.8949
		ARCNN [9]	52.27	0.8969		ARCNN [9]	52.2	0.9012
		Proposed-4	53.31	0.9194		Proposed-4	53.38	0.9315
		Proposed-8	**54.06**	**0.9309**		Proposed-8	**54.42**	**0.9367**

C. *Discussion of Fusion Order*

In Table I, the "Late-4" represents the late fusion network with 5 convolutional layers, and each layer has 64 filters with kernel size 3 × 3. However, in depth it has 4 layers, which is the same as the early fusion network and ARCNN. We find that the early fusion 4-layer architecture in Fig. 1 has slightly better performance than the late fusion architecture in Fig. 2. In the late fusion network, there are two groups of independent filters for the LQ and virtual view images



(a) GT (b) LQ image (c) Virtual view image (d) L-Conv1 LQ (e) L-Conv1 virtual (f) L-Conv2

(g) L-Conv3 (h) L-Conv4 (i) E-Conv1 (j) E-Conv2 (k) E-Conv3 (l) E-Conv4

Figure 4. The feature map comparison between the late fusion and early fusion networks. (a)-(c) are the uncompressed image, compressed LQ image and corresponding virtual view image; (d)-(h) are the feature maps obtained from the corresponding layers of the late fusion network, and (i) -(l) are the feature maps obtained from the corresponding layers of the early fusion network.

at the first convolutional layer, respectively. Therefore, the occluded regions in the virtual view image have a larger effect on its corresponding feature maps (Fig. 4 (e)), which later on are concatenated with the feature maps generated from the LQ image (Fig. 4 (d)) and fed into the second layer. From the feature maps comparison in Fig. 4, we can notice that after the activation layer, the second layer feature maps of the late fusion architecture has less non-zero feature maps (Fig. 4 (f)). Moreover, the majority of these features contain “hole features”, which have negative effects on the final reconstructed images. These negative effects, however, are significantly suppressed after the third layer (Fig. 4 (g)). Finally, the late fusion network obtains the similar residual map (Fig. 4 (h)) as the early fusion one (Fig. 4 (l)). However, due to obtaining less valid feature maps in later layers, the late fusion performance is still worse than that of the early fusion.

IV. Conclusion

In this paper, an early fusion based end-to-end CNN is proposed to enhance the quality of the LQ image by exploiting the HQ image in quality-asymmetric stereo images. Through the proposed network architecture, HF information contained in the virtual view image is extracted and fused with the LQ image in order to recover the lost information in the LQ image during compression. Extensive experiments demonstrate our method outperforms the existing methods on the testing image pairs and MPEG images.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (NSFC) (61771321, 61701313, 6171101606, 61472257); Natural Science Foundation of Shenzhen (KQJSCX20170327151357330, JCYJ20170818091621856, JCYJ20170302145906843); TUM University Foundation Fellowship (TUFF) and China Postdoctoral Science Foundation Grants (2017M622778).

References

[1] C. Fehn, “Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3d-tv,” *Electronic Imaging 2004*, pp. 93–104, 2004.

[2] B. Randolph, “Threshold conditions for binocular rivalry,” *Journal of Experimental Psychology-human Perception and Performance*, vol. 3, pp. 251–257, 1977.

[3] P. Aflaki, M.M. Hannuksela, J. Ha?kkinen, P. Lindroos, and M. Gabbouj, “Subjective study on compressed asymmetric stereoscopic video,” in *ICIP*, Sept. 2010, pp. 4021 –4024.

[4] G. K. Wallace, “The jpeg still picture compression standard,” *IEEE TCE*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

[5] G. Saygili, C.G. Gurler, and A M. Tekalp, “Evaluation of asymmetric stereo video coding and rate scaling for adaptive 3d video streaming,” *IEEE TB*, vol. 57, no. 2, pp. 593–601, 2011.

[6] P. Aflaki, M.M Hannuksela, and M. Gabbouj, “Subjective quality assessment of asymmetric stereoscopic 3d video,” *Signal, Image and Video Processing*, pp. 1–15, 2013.

[7] A. Foi, V. Katkovnik, and K. Egiazarian, “Pointwise shape-adaptive dct for high-quality denoising and deblocking of grayscale and color images,” *IEEE TIP*, vol. 16, no. 5, pp. 1395–1411, May. 2007.

[8] J. Jancsary, S. Nowozin, and C. Rother, “Loss-specific training of non-parametric image restoration models: A new state of the art,” *ECCV*, pp. 112–125, Oct. 2012.

[9] C. Dong, Y. Deng, C. Change Loy, and X. Tang, “Compression artifacts reduction by a deep convolutional network,” *IEEE ICCV*, pp. 576–584, Dec. 2015.

[10] K. Yu, C. Dong, C.C.Loy, and X.Tang, “Deep convolution networks for compression artifacts reduction,” *arXiv preprint arXiv:1608.02778*, 2016.

[11] Z. Jin, L. Luo, Y. Tang, W. Zou, and X. Li, “A cnn cascade for quality enhancement of compressed depth images,” *IEEE VCIP*, pp. 1–4, Dec. 2017.

[12] Z. Jin, T. Tillo, and L. Luo, “Quality enhancement of quality-asymmetric multiview plus depth video by using virtual view,” *ICME*, pp. 1–6, Jul. 2015.

[13] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, “High-resolution stereo datasets with subpixel-accurate ground truth,” *GCPR*, pp. 31–42, Sept. 2014.

[14] Matthew D Zeiler, “Adadelta: an adaptive learning rate method,” *arXiv preprint arXiv:1212.5701*, 2012.

[15] Y. H. Lin and J. L. Wu, “Quality assessment of stereoscopic 3d image compression by binocular integration behaviors,” *IEEE TIP*, vol. 23, no. 4, pp. 1527–1542, Apr. 2014.