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
The proposed CNN consists of images as compared to the state-of-the-art approaches. Our approach can efficiently enhance the quality of the LQ and reconstruction accuracy. Extensive experiments indicate Residual learning is adopted to boost up the training process and information fusion happens at the first convolutional layer.

The LQ image as the two inputs of the proposed network. The map at the viewpoint of the LQ image is combined with the view image generated from the HQ image and the depth. 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. 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 I_{LQ} and the virtual view image 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:

\[ F_1(I_{LQ}, I_V) = ReLU(W_{1,1} \ast I_{LQ} + W_{1,2} \ast I_V + B_1) \] (1)

where W and B represent the weights and biases, respectively; "\ast" 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 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:

\[ F_i(I) = ReLU(W_i \ast F_{i-1}(I) + B_i) \] (2)

where i indicates the i-th convolutional layer; F_{i}(I) and F_{i-1}(I) are the outputs of the i-th and i−1-th convolutional layer, respectively; W_{i} and B_{i} represent the weights and biases for the i-th layer, respectively.

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

\[ F_n(I) = W_n \ast F_{n-1}(I) + B_n \] (3)
Moreover, the final enhanced LQ image can be expressed as:

\[ F(I_{LQ}, I_V) = F_u(I) + I_{LQ} \]  

(4)

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

\[ \text{Loss}(\Theta) = \frac{1}{m} \sum_{i=1}^{m} \| F(I_{LQ}, I_V; \Theta) - I_{GT} \|^2 \]  

(5)

where \( \Theta \) 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.

#### A. Implementation Details

To train the proposed network, we adopt the Middlebury 2014 stereo image dataset [13] containing 18 images of size \( 2964 \times 1988 \) as our training data. The remains 5 stereo images and 7 first frames of the MPEG multiview 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 \( 24 \times 24 \) 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 \( Q_F = 3, 5, 7 \) 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
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 Middlebury 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.

![Figure 3. Visual comparison results that achieved with the proposed CNN and other state-of-the-art approaches on MPEG image “Shark”.](image-url)

### Table II

<table>
<thead>
<tr>
<th>QF</th>
<th>Algo.</th>
<th>FI-PSNR</th>
<th>FI-SSIM</th>
<th>Algorithms</th>
<th>FI-PSNR</th>
<th>FI-SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>JPEG [4]</td>
<td>45.77</td>
<td>0.8231</td>
<td>JPEG [4]</td>
<td>45.77</td>
<td>0.8231</td>
</tr>
<tr>
<td></td>
<td>SA-DCT [7]</td>
<td>45.98</td>
<td>0.8315</td>
<td>SA-DCT [7]</td>
<td>45.98</td>
<td>0.8315</td>
</tr>
<tr>
<td></td>
<td>ARCNN [9]</td>
<td>46.47</td>
<td>0.8491</td>
<td>ARCNN [9]</td>
<td>46.47</td>
<td>0.8491</td>
</tr>
<tr>
<td></td>
<td>Proposed-4</td>
<td>46.27</td>
<td>0.9144</td>
<td>Proposed-4</td>
<td>46.27</td>
<td>0.9144</td>
</tr>
<tr>
<td></td>
<td>Proposed-8</td>
<td>46.37</td>
<td>0.9367</td>
<td>Proposed-8</td>
<td>46.37</td>
<td>0.9367</td>
</tr>
<tr>
<td>20</td>
<td>JPEG</td>
<td>52.54</td>
<td>0.8669</td>
<td>JPEG</td>
<td>52.54</td>
<td>0.8669</td>
</tr>
<tr>
<td></td>
<td>SA-DCT [7]</td>
<td>52.82</td>
<td>0.9049</td>
<td>SA-DCT [7]</td>
<td>52.82</td>
<td>0.9049</td>
</tr>
<tr>
<td></td>
<td>ARCNN [9]</td>
<td>52.2</td>
<td>0.9002</td>
<td>ARCNN [9]</td>
<td>52.2</td>
<td>0.9002</td>
</tr>
<tr>
<td></td>
<td>Proposed-4</td>
<td>54.06</td>
<td>0.9369</td>
<td>Proposed-4</td>
<td>54.06</td>
<td>0.9369</td>
</tr>
<tr>
<td></td>
<td>Proposed-8</td>
<td>54.42</td>
<td>0.9367</td>
<td>Proposed-8</td>
<td>54.42</td>
<td>0.9367</td>
</tr>
</tbody>
</table>

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 $3 \times 3$. 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.
Fusion performance is still worse than that of the early fusion. Due to obtaining less valid feature maps in later layers, the late map (Fig. 4 (h)) as the early fusion one (Fig. 4 (l)). However, are significantly suppressed after the third layer (Fig. 4 (g)). Final reconstructed images. These negative effects, however, contain “hole features”, which have negative effects on the maps (Fig. 4 (f)). Moreover, the majority of these features maps of the late fusion architecture has less non-zero feature (Fig. 4 (f)). Notice that after the activation layer, the second layer feature 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