Deep Stereo Image Compression via Bi-directional Coding

Jianjun Lei[1] Xiangrui Liu[1] Bo Peng[1*] Dengchao Jin[1] Wanqing Li[2] Jingxiao Gu[3]
[1] Tianjin University [2] University of Wollongong [3] CalmCar Vehicle Vision System
{jjlei,xr_liu,bpeng,jdc3159761141}@tju.edu.cn, wanqing@uow.edu.au, jingxiao.gu@calmcar.com

Abstract

Existing learning-based stereo compression methods usually adopt a unidirectional approach to encoding one image independently and the other image conditioned upon the first. This paper proposes a novel bi-directional coding-based end-to-end stereo image compression network (BCSIC-Net). BCSIC-Net consists of a novel bi-directional contextual transform module which performs nonlinear transform conditioned upon the inter-view context in a latent space to reduce inter-view redundancy, and a bi-directional conditional entropy model that employs inter-view correspondence as a conditional prior to improve coding efficiency. Experimental results on the InStereo2K and KITTI datasets demonstrate that the proposed BCSIC-Net can effectively reduce the inter-view redundancy and outperforms state-of-the-art methods.

1. Introduction

With the rapid development of stereoscopic imaging technologies, stereo images are widely used in many applications, such as augmented reality, autonomous driving, and robot navigation [23, 11, 31, 21]. Accordingly, several methods [8, 14, 15, 18, 19, 24, 13] have been developed and improved to compress stereo images. Different from single image compression, compression of stereo images needs to reduce the inter-view redundancy in addition to the intra-view redundancy.

Traditionally, stereo image compression methods employ inter-view prediction, such as disparity compensation prediction (DCP) [8, 15, 18, 19], to deal with the inter-view redundancy. For instance, when compressing the right image, DCP estimates the disparities between the right image and the reconstructed left image, then derives the prediction of the right image via disparity compensation. The disparities and the residues between the actual and predicted right image are encoded. However, the hand-crafted prediction methods struggle to cope with the intricate inter-view correlations for complex scenes, and thus the residues and the bits often remain large.

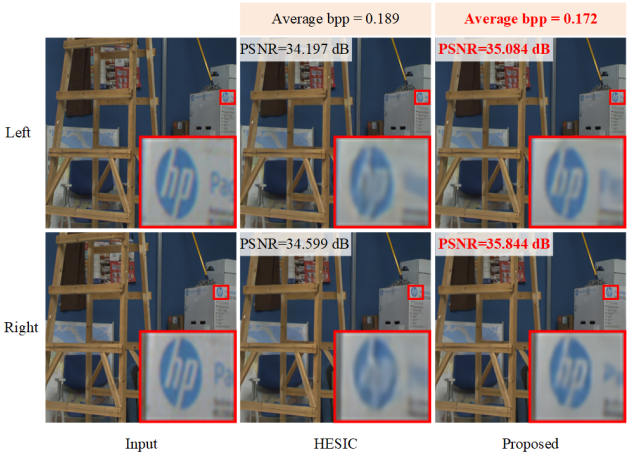

Figure 1. Illustration of compression performance achieved by the proposed method and the state-of-the-art method HESIC [13].

With the development of deep learning, end-to-end single image compression [2, 3, 27, 20, 22, 26, 10, 12, 25, 17, 16] has achieved promising progress. These works have provided insights and methodologies for stereo image compression, and some frameworks [24, 13] following the deep learning paradigm have been preliminarily studied. Specifically, they adopt deep learning-based single image compression networks and inter-view prediction networks to reduce the intra-view and inter-view redundancy, respectively. The existing methods mainly employ a unidirectional coding mechanism to reduce the inter-view redundancy, *i.e.*, the reconstruction [8, 18, 19, 13] or the latent representation [24] of the left image is propagated as a context to the right-view coding branch. Therefore, they follow on the strictly sequential coding order that the left image is first encoded and then the right image or vice versa. Such a unidirectional framework, on the one hand, is not always effective to reduce the inter-view redundancy. On the other hand, it is difficult to extend the framework into a bi-directional coding framework that is expected to be more effective in reducing the inter-view redundancy, hence, saving the bitrate.

*Corresponding author.

To address this issue, this paper proposes an end-to-end stereo image compression network based on bi-directional coding (BCSIC-Net). The main idea of the proposed method is to eliminate the limitation of sequential coding by designing a novel inter-view context dependency, *i.e.*, extending the unidirectional coding mechanism to a bi-directional one. Thus, a bi-directional contextual transform module (Bi-CTM) and a bi-directional conditional entropy model (Bi-CEM) are proposed. As shown in Figure. 1, compared with the state-of-the-art unidirectional method [13], the proposed BCSIC-Net can achieve higher reconstruction quality with lower bit-consuming.

The major contributions of this paper are summarized as follows.

1) A novel end-to-end stereo image compression network based on bi-directional coding (BCSIC-Net) is proposed to improve the performance of stereo image compression by effectively exploiting the inter-view correlation.

2) A bi-directional contextual transform module (Bi-CTM) that performs nonlinear transform conditioned on the inter-view context is presented to effectively reduce the redundancy between stereo views.

3) A bi-directional conditional entropy model (Bi-CEM) is developed to improve the efficiency of entropy coding by exploiting the inter-view correspondence as a conditional prior.

4) Experimental results on popular benchmark datasets show that the proposed method achieves the state-of-the-art coding performance.

The rest of this paper is organized as follows. Section II summarizes the related works of the single image compression and stereo image compression. The proposed method is described in Section III, followed by experimental results and analysis in Section IV. Section V concludes the paper.

2. Related Works

This section briefly reviews the recent methods for single image compression and stereo image compression.

2.1. Single Image Compression

Traditional image compression methods typically consist of hand-crafted transform of an image into compact coefficients, quantization of the coefficients, and entropy coding of the quantized coefficients [33, 30]. Furthermore, hybrid coding methods employed intra prediction to reduce spatial redundancy [6, 32, 9].

In the era of deep learning, various end-to-end image compression methods have been investigated [2, 3, 27, 20, 22, 26, 10, 12, 25, 17, 16]. These methods employ neural networks to nonlinearly transform an image to a compact latent representation. An entropy model is then used to estimate the probability distribution of the latent representation for entropy coding. In decoder, the reconstructed image is generated from the latent representation. The coding performance of the end-to-end framework mainly depends on how well the nonlinear transform and the entropy model can be learned [4].

Recently, several studies have been done on nonlinear transform to improve compact representation. Ballé *et al.* [2] proposed an end-to-end image compression method, which utilizes the generalized divisive normalization (GDN) and inverse GDN (IGDN) [1] to strengthen the nonlinearity of the transform. Cheng *et al.* [12] incorporated an attention mechanism with the nonlinear transform to derive a latent representation. Ma *et al.* [25] proposed a wavelet-like invertible transform that avoids information loss in the nonlinear transform.

Further studies have also been reported to improve the efficacy of the entropy model. For instance, Ballé *et al.* [3] proposed an entropy model conditioned on the side information, namely the hyperprior, in which the Gaussian distribution is used to model the probability distribution of latent representation and its parameters are determined by the hyperprior. Minnen *et al.* [27] proposed an autoregressive context as a supplementary prior to promote the efficiency of the conditional entropy model in [3]. Inspired by [27], Lee *et al.* [20] proposed an adaptive context entropy model. Chen *et al.* [10] developed an autoregressive context model based on a 3D-CNN to capture both spatial and channel correlations in the latent representation. Additionally, several methods have been reported to reduce the computational complexity of entropy model. Hu *et al.* [17] obtained considerable time savings by replacing the autoregressive model with a coarse-to-fine hyperprior structure. He *et al.* [16] suggested a parallel variation of the autoregressive context model named the checkerboard context model to accelerate the decoding process.

Despite the progress on end-to-end single image compression and the fact that such compression can be applied to the left and right views of stereo images, a mechanism is required to reduce the inter-view redundancy in stereo images in order to further improve coding efficiency.

2.2. Stereo Image Compression

In stereo image compression, inter-view redundancy needs to be reduced in addition to intra-view redundancy. Typically, stereo image compression follows a unidirectional coding approach, *i.e.*, the left image is encoded independently, then the right image is encoded conditioned on the context provided by the compressed left image.

Traditionally, inter-view prediction, such as DCP, is performed to reduce the redundancy between the left and right images. Boulgouris *et al.* [8] proposed a stereo image compression framework based on DCP. Specifically, the left image is first independently encoded, and then the reconstructed left image is specified as the reference image of

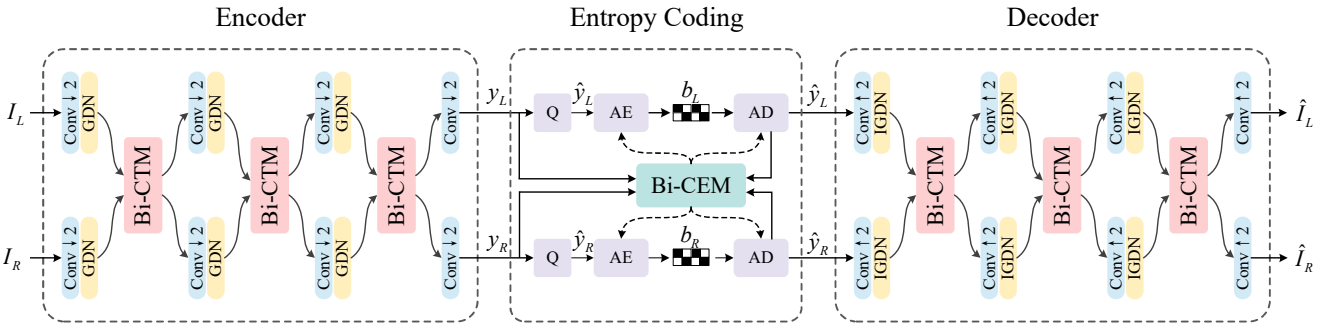

Figure 2. Architecture of the proposed BCSIC-Net.

the right image. When coding the right image, a prediction of the right image is generated based on the reference image using DCP, and only the estimated disparities and prediction residues are compressed. Kaaniche *et al.* [18] incorporated the vector lifting scheme with DCP to effectively compress the prediction residues. Kadaikar *et al.* [19] proposed a variable size-block stereo image compression method, which combines the variable block-size coding strategy with DCP.

Recently, researchers have proposed some end-to-end methods for stereo image compression [24, 13]. Liu *et al.* [24] proposed deep stereo image compression (DSIC), where a parametric skip function is designed to exploit the inter-view correlations for improving coding performance. Specifically, the parametric skip function warps the left-view features and latent representation into the right-view coding branch to provide inter-view information. Additionally, a conditional entropy model is employed, in which the left latent representation serves as a prior for the right-view probability distribution. Deng *et al.* [13] developed deep homography for efficient stereo image compression (HESIC), in which a homography matrix is introduced to produce the right prediction by warping the reconstructed left image. Following that, the right-view coding branch only compresses the homography matrix and the prediction residues.

However, these unidirectional coding methods sequentially encode the left and right views and are hard to extend to a bi-directional coding scheme.

3. Proposed Method

3.1. Architecture of BCSIC-Net

As illustrated in Figure. 2, the proposed BCSIC-Net consists of three components, including encoder, entropy coding, and decoder. Specifically, a pair of left and right images $\{I_L, I_R\}$ are transformed to latent representation $\{y_L, y_R\}$ by the encoder. Inside the encoder, the proposed Bi-CTM is inserted after each GDN [1] layer to reduce the redundancy between the left and right features. Owing to the employment of Bi-CTM at multiple feature levels in the encoder,

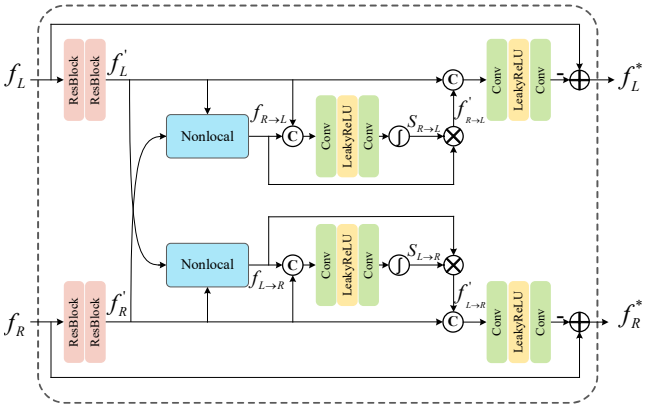

Figure 3. Illustration of the bi-directional contextual transform module.

BCSIC-Net is capable of improving the coding performance by sufficiently reducing the inter-view redundancy at multiple feature scales.

During the entropy coding, quantization is firstly applied to $\{y_L, y_R\}$. The proposed Bi-CEM is then utilized to provide probability estimates for the quantized latent representation $\{\hat{y}_L, \hat{y}_R\}$. Inside the Bi-CEM, inter-view correspondence is effectively exploited to improve probability estimates. Finally, an arithmetic encoding (AE) is performed to compress $\{\hat{y}_L, \hat{y}_R\}$ into bitstream $\{b_L, b_R\}$, and the arithmetic decoding (AD) is used to retrieve $\{\hat{y}_L, \hat{y}_R\}$ from $\{b_L, b_R\}$.

The decoder is utilized to transform the quantized latent representation $\{\hat{y}_L, \hat{y}_R\}$ to the left and right images $\{\hat{I}_L, \hat{I}_R\}$. It is symmetric to the encoder, and the proposed Bi-CTM is also inserted after each IGDN [1] layer.

3.2. Bi-directional Contextual Transform Module

Due to the high degree of similarity in contents, there exists significant inter-view redundancy between the left and right features. Therefore, the Bi-CTM is proposed to effectively reduce the inter-view redundancy by performing a nonlinear transform conditioned upon the inter-view context in a latent space.

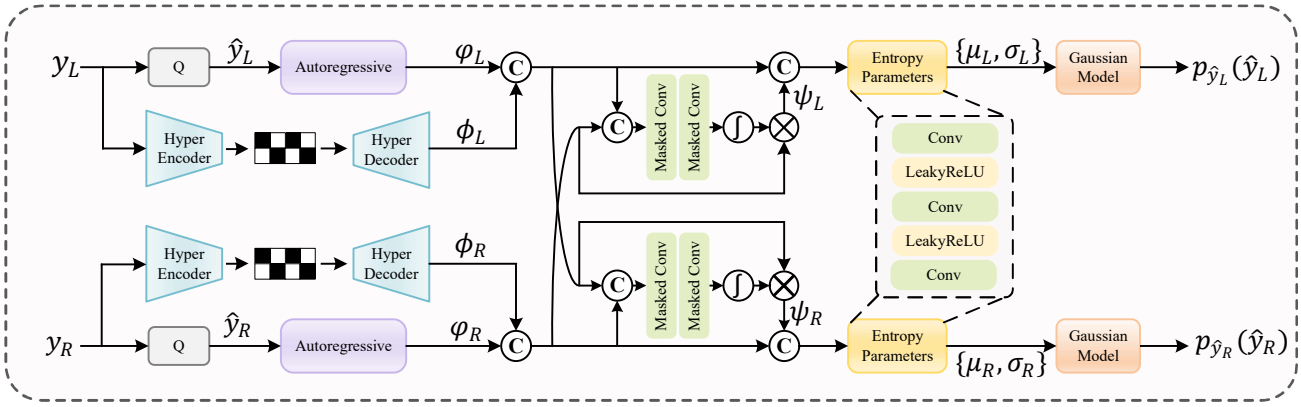

Figure 4. Illustration of the bi-directional conditional entropy model.

As depicted in Figure. 3, the proposed Bi-CTM takes the left and right features $\{f_L, f_R\}$ as inputs and outputs compact features $\{f_L^*, f_R^*\}$. At the start of the proposed Bi-CTM, two residual blocks are separately applied to $\{f_L, f_R\}$ to generate representative features $\{f_L^{'}, f_R^{'}\}$. Inside the proposed Bi-CTM, there are two parallel paths. In the left-view path, a two-stage mapping function is developed to generate context for f_L. Specifically, in the first stage, the preliminary context $f_{R\to L}$ is acquired by mapping the right feature $f_R^{'}$ to the left view,

$$f_{R\to L} = F_L(f_R^{'}, f_L^{'}), \tag{1}$$

where $F_L(\cdot)$ denotes the mapping function implemented by a simplified nonlocal block [29]. In the subsequent stage, the refined context $f_{R\to L}^{'}$ is obtained by screening $f_{R\to L}$ according to $f_L^{'}$,

$$\begin{aligned} f_{R\to L}^{'} &= S_{R\to L} * f_{R\to L} \\ with \quad S_{R\to L} &= \sigma(h_L(f_{R\to L} \oplus f_L^{'})), \end{aligned} \tag{2}$$

where $h_L(\cdot)$ is composed of two consecutive convolution layers, $\sigma(\cdot)$ indicates the sigmoid function, and $\oplus$ denotes the channel-wise concatenation. Finally, f_L is transformed to a more compact feature f_L^* conditioned on the inter-view context $f_{R\to L}^{'}$,

$$f_L^* = f_L - g_L(f_L^{'} \oplus f_{R\to L}^{'}), \tag{3}$$

where $g_L(\cdot)$ is composed of two consecutive convolution layers. Similar process is applied to the right feature f_R to reduce the inter-view redundancy in f_R and yield the compact feature f_R^*.

The proposed Bi-CTM explores the inter-view contexts $\{f_{R\to L}^{'}, f_{L\to R}^{'}\}$ with respect to left and right features from the right and left views, respectively. Compact features $\{f_L^*, f_R^*\}$ are then extracted conditioned on $\{f_{R\to L}^{'}, f_{L\to R}^{'}\}$. Compared with the existing unidirectional mechanism, the proposed Bi-CTM provides a novel conditional dependency, *i.e.*, the bi-directional context for effective reduction of inter-view redundancy.

3.3. Bi-directional Conditional Entropy Model

Appropriate conditional dependencies are critical for the estimation of probabilities in a conditional entropy model. Therefore, the Bi-CEM is developed to exploit the inter-view correspondence as an additional prior for the left and right latent representation, namely the inter-view prior.

As shown in Figure. 4, Bi-CEM estimates the probability distribution $\{p_{\hat{y}_L}(\hat{y}_L), p_{\hat{y}_R}(\hat{y}_R)\}$ for the quantized latent representation $\{\hat{y}_L, \hat{y}_R\}$. Inter-view prior is utilized to provide conditional dependencies for the input latent representation and integrated into the autoregressive entropy model [27], *i.e.*,

$$\begin{aligned} p_{\hat{y}_L}(\hat{y}_L) &= \prod_i p_{\hat{y}_L}(\hat{y}_L^i | \varphi_L, \phi_L^{<i}, \psi_L^{<i}), \\ p_{\hat{y}_R}(\hat{y}_R) &= \prod_j p_{\hat{y}_R}(\hat{y}_R^j | \varphi_R, \phi_R^{<j}, \psi_R^{<j}), \end{aligned} \tag{4}$$

where $\hat{y}_L^i$ represents the i^{th} elements of $\hat{y}_L$, and $\hat{y}_R^j$ represents the j^{th} elements of $\hat{y}_R$. The priors $\{\varphi_L, \phi_L^{<i}, \psi_L^{<i}\}$ denote the hyperprior, the autoregressive prior, and the proposed inter-view prior for $\hat{y}_L^i$, respectively. Similarly, $\{\varphi_R, \phi_R^{<j}, \psi_R^{<j}\}$ indicate the priors for $\hat{y}_R^j$.

Since the hyperprior and autoregressive prior are practical representations of the image content, the inter-view prior is generated based on the two priors. In particular, for the i^{th} element of the left latent representation, the inter-view prior $\psi_L^{<i}$ can be calculated as follows:

$$\begin{aligned} \psi_L^{<i} &= \sigma(u_L(\pi_L^{<i} \oplus \pi_R^{<i})) * \pi_R^{<i} \\ with \quad \pi_L^{<i} &= \varphi_L \oplus \phi_L^{<i} \; and \; \pi_R^{<i} = \varphi_R \oplus \phi_R^{<i}, \end{aligned} \tag{5}$$

where $u_L(\cdot)$ is composed of two masked convolution layers. The inter-view prior of the j^{th} right elements $\psi_R^{<j}$

Table 1. BD-PSNR and BD-rate comparisons on the InStereo2K dataset.

Methods	BD-PSNR (dB)	BD-rate (%)
Ballé (ICLR'17) [2]	-0.489	14.195
BPG [6]	-0.501	14.162
HEVC/H.265 [32]	-0.005	-11.342
Lee (ICLR'19) [20]	0.192	-8.167
Hu (AAAI'20) [17]	0.169	-4.415
DSIC (ICCV'19) [24]	0.238	-7.062
HESIC (CVPR'21) [13]	1.373	-38.809
Proposed	**1.778**	**-45.745**

Table 2. BD-PSNR and BD-rate comparisons on the KITTI dataset.

Method	BD-PSNR (dB)	BD-rate (%)
Ballé (ICLR'17) [2]	-0.311	16.750
BPG [6]	-1.418	105.068
HEVC/H.265 [32]	-1.367	105.804
Lee (ICLR'19) [20]	-0.897	55.633
Hu (AAAI'20) [17]	-0.677	41.406
DSIC (ICCV'19) [24]	0.005	-4.027
HESIC (CVPR'21) [13]	0.920	-28.836
Proposed	**1.518**	**-39.068**

can be calculated in the same way. Additionally, the Gaussian conditional model is used to parametrically model the $\{p_{\hat{y}_L}(\hat{y}_L), p_{\hat{y}_R}(\hat{y}_R)\}$, *i.e.*,

$$\begin{aligned} p_{\hat{y}_L}(\hat{y}_L^i|\varphi_L, \phi_L^{<i}, \psi_L^{<i}) &\sim \mathcal{N}(\mu_L^i, \sigma_L^i), \\ p_{\hat{y}_R}(\hat{y}_R^j|\varphi_R, \phi_R^{<j}, \psi_R^{<j}) &\sim \mathcal{N}(\mu_R^j, \sigma_R^j), \end{aligned} \tag{6}$$

The Gaussian parameters are estimated by the priors,

$$\begin{aligned} \mu_L^i, \sigma_L^i &= v_L(\varphi_L, \phi_L^{<i}, \psi_L^{<i}), \\ \mu_R^j, \sigma_R^j &= v_R(\varphi_R, \phi_R^{<j}, \psi_R^{<j}), \end{aligned} \tag{7}$$

where $v_L(\cdot)$ and $v_R(\cdot)$ indicate the estimators for the left and right views, respectively.

The proposed Bi-CEM utilizes the inter-view priors $\{\psi_L, \psi_R\}$ to improve the estimates of probabilities $\{p_{\hat{y}_L}(\hat{y}_L), p_{\hat{y}_R}(\hat{y}_R)\}$. Compared with the existing conditional entropy models for stereo image compression [24, 13], the suggested entropy model generalizes the unidirectional probability dependency to a bi-directional one, in which the latent representation $\{\hat{y}_L, \hat{y}_R\}$ can be conditioned on the inter-view correspondence concurrently.

3.4. Implementation

The proposed BCSIC-Net is trained using the same rate-distortion function $\mathcal{L}$ as in the previous works [24, 13],

$$\mathcal{L} = R_L + R_R + \lambda(D_L + D_R), \tag{8}$$

where $\{R_L, R_R\}$ denote the bitrate calculated from the estimated probabilities $\{p_{\hat{y}_L}(\hat{y}_L), p_{\hat{y}_R}(\hat{y}_R)\}$ respectively, and $\{D_L, D_R\}$ represent the reconstruction distortion calculated by mean square error metric. λ is a hyper parameter to control the rate-distortion trade-off. Specifically, λ is set to 128, 256, 512, 1024, 2048, 4096 and 8192 to meet various bitrates.

In order to conduct a fair comparison with HESIC that has a post-processing network, a post-processing network consisting of residual blocks is utilized. Specifically, two identical branches consisting of 16 residual blocks are applied to enhance the quality of the reconstructed left and right images, respectively.

4. Experiments

4.1. Experimental Settings

To ensure the fairness of experiments, the same datasets provided by [13], *i.e.* InStereo2K and KITTI, are adopted to train the network and conduct the tests. The proposed method is implemented based on compreessAI [5] and PyTorch [28]. Adam optimizer is adopted to train the network for 400 epochs with initial learning rate of 0.0001 and parameters $\beta_1 = 0.9, \beta_2 = 0.999$. The learning rate decays by half per 100 epochs. All experiments are conducted on a PC with GeForce GTX 1080Ti GPU and Intel i7-8700K processor @3.70 GHz.

4.2. Results and Analysis

Objective evaluation. To demonstrate the efficacy of the proposed BCSIC-Net, quantitative evaluations are conducted to compare the proposed method with the traditional compression methods [6, 32] as well as several end-to-end compression methods, including methods designed for single image [2, 20, 17] and methods designed for stereo image [24, 13]. Since the experimental settings are same as those in [13], the results of [2, 20, 17, 24, 13] are obtained from [13]. In the experiments, the widely employed BD-rate and BD-PSNR [7] are adopted as the objective metrics to evaluate the coding performance, where the Ballé *et al.* [3] is set as the baseline to compute the BD-rate and BD-PSNR. As shown in Table 1, on the InStereo2K dataset, the proposed BCSIC-Net obtains 45.745% BD-rate reduction and 1.778 dB BD-PSNR increase. HESIC [13] obtains 38.809% BD-rate reduction and 1.373 dB BD-PSNR increase. Compared with HESIC [13], the proposed BCSIC-Net further achieves 6.936% BD-rate reduction and 0.405 dB BD-PSNR increase. In addition, experiments on the KITTI dataset are also conducted to verify the effectiveness of the proposed

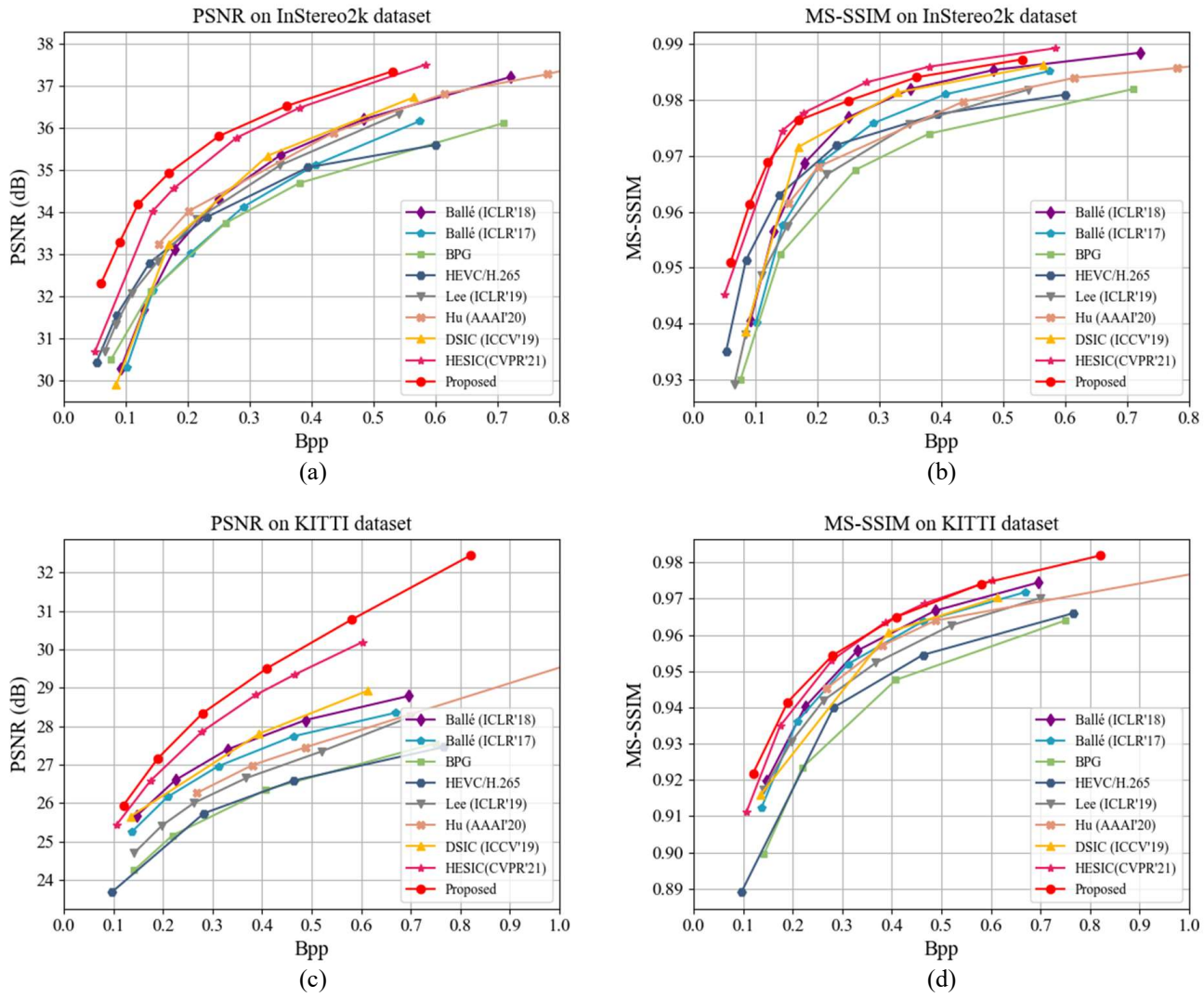

Figure 5. Rate-distortion curves in terms of PSNR and MS-SSIM on the InStereo2K and KITTI datasets. (a) Comparison of PSNR on the InStereo2K dataset, (b) Comparison of MS-SSIM on the InStereo2K dataset, (c) Comparison of PSNR on the KITTI dataset, (d) Comparison of MS-SSIM on the KITTI dataset.

BCSIC-Net, and results are shown in Table 2. It can be observed that, compared with with HESIC [13], the proposed BCSIC-Net has significant advantages. Specifically, HESIC [13] obtains 25.967% BD-rate reduction and 0.920 dB BD-PSNR increase, while BCSIC-Net obtains 39.068% BD-rate reduction and 1.518 dB BD-PSNR increase. The experimental results in Table 1 and Table 2 show that the proposed BCSIC-Net achieves the highest rate-distortion performance on both the InStereo2K and KITTI datasets, which indicates the advantage of the proposed BCSIC-Net.

In addition, rate-distortion curves in terms of the PSNR and the multi-scale structural similarity (MS-SSIM) metrics are shown in Figure. 5. As for the PSNR metric, the proposed BCSIC-Net yields better performance than other comparison methods on both the InStereo2K and KITTI datasets, which demonstrates the effectiveness of the proposed BCSIC-Net. Regarding the MS-SSIM metric, the results in Figure. 5 demonstrate that the performance of the proposed BCSIC-Net is competitive with other methods.

Table 3. Performance comparison in ablation study.

	BD-PSNR (dB)	BD-rate (%)
w/o Bi-CTM	0.950	-29.174
w/o Bi-CEM	1.187	-32.837
Proposed	1.778	-45.745

Visual evaluation. Four reconstructed images from the InStereo2K test set are shown in Figure. 6, and some details are zoomed in. It can be observed that the reconstructed images obtained by the proposed BCSIC-Net have clearer

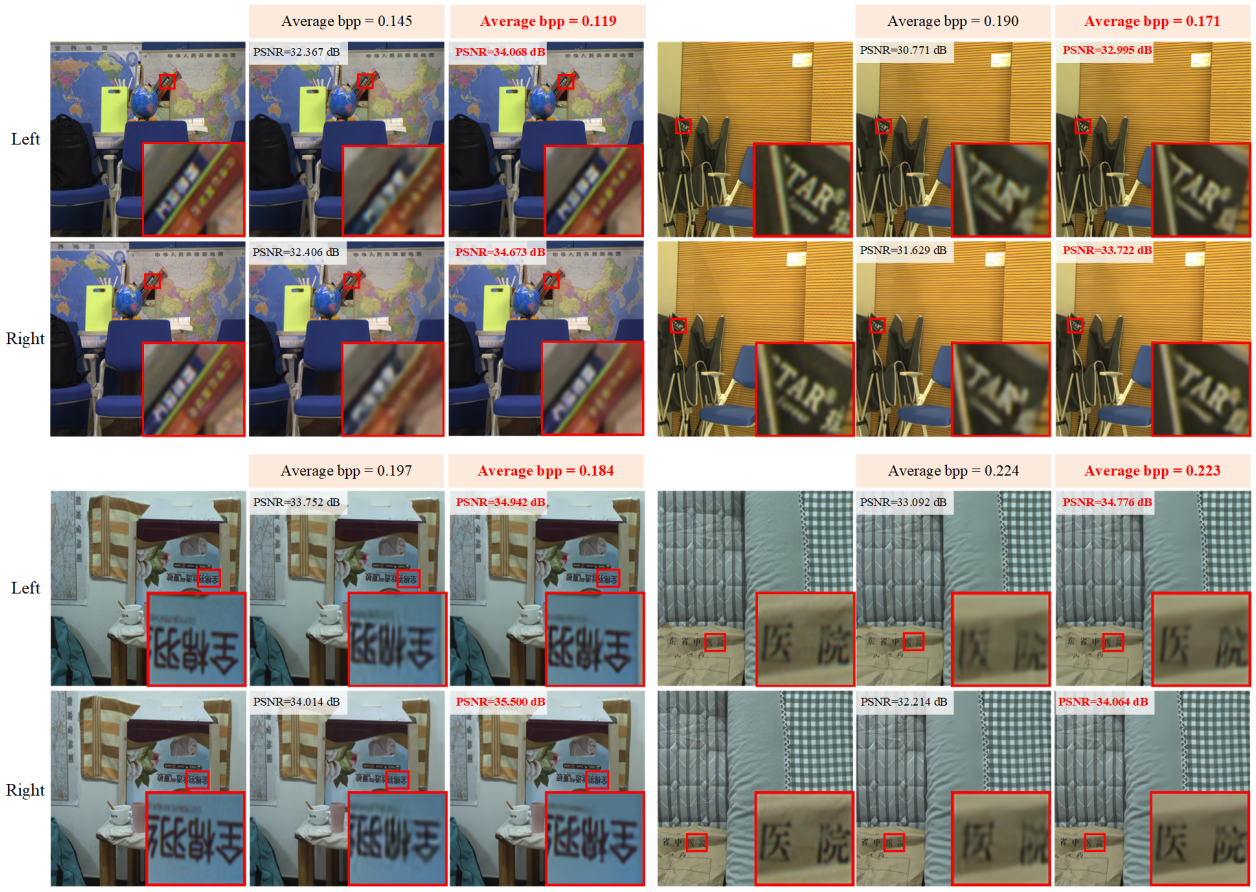

Figure 6. Visual comparison of the reconstructed left and right images from the proposed BCSIC-Net and the HESIC [13]. For each comparison case, the first column is the input images, the second column is the images reconstructed by the HESIC [13], and the third column is the images reconstructed by the proposed BCSIC-Net.

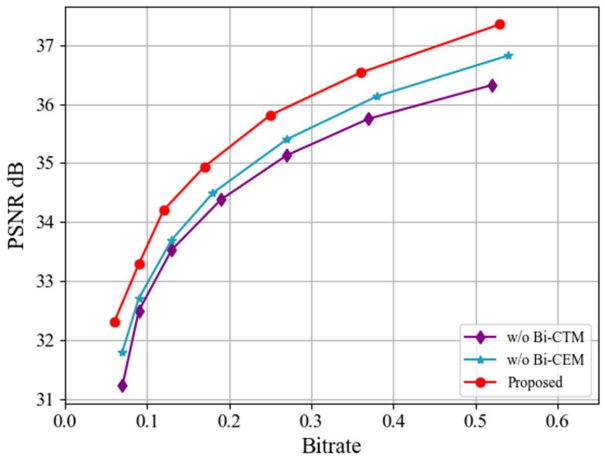

Figure 7. Ablation study on the InStereo2K dataset.

texture and edges as compared to HESIC [13].

4.3. Ablation Study

1) Effectiveness of the Bi-directional Contextual Transform Module. The Bi-CTM is designed to effectively reduce the redundancy between the left and right views by performing a nonlinear transform conditioned on the inter-view context. To evaluate the contribution of the bi-directional context in Bi-CTM, f_R is no longer fed to the left branch and the up nonlocal block as well as subsequent convolution layers are removed. That is, f_L is processed without concatenating any information about f_R. The experimental results are shown in Figure. 7 and Table 3, denoted as "w/o Bi-CTM", while the method in [3] is assigned as the anchor for the calculation of the BD-rate and BD-PSNR. It can be observed that, without the bi-directional context in the Bi-CTM, the BD-PSNR performance decreases from 1.778 dB to 0.950 dB and the BD-rate performance decreases from 45.745% to 29.174%. This is because the left feature is disabled to exploit the inter-view correlations, resulting in an ineffective reduction of the inter-view redundancy.

To further illustrate the effectiveness of the Bi-CTM, the latent representation of the BCSIC-Net and the BCSIC-Net without Bi-CTM are visualized in Figure 8. It can be observed that the left latent representation generated by the BCSIC-Net is more compact than that of the BCSIC-Net without Bi-CTM, while the right latent representation gen-

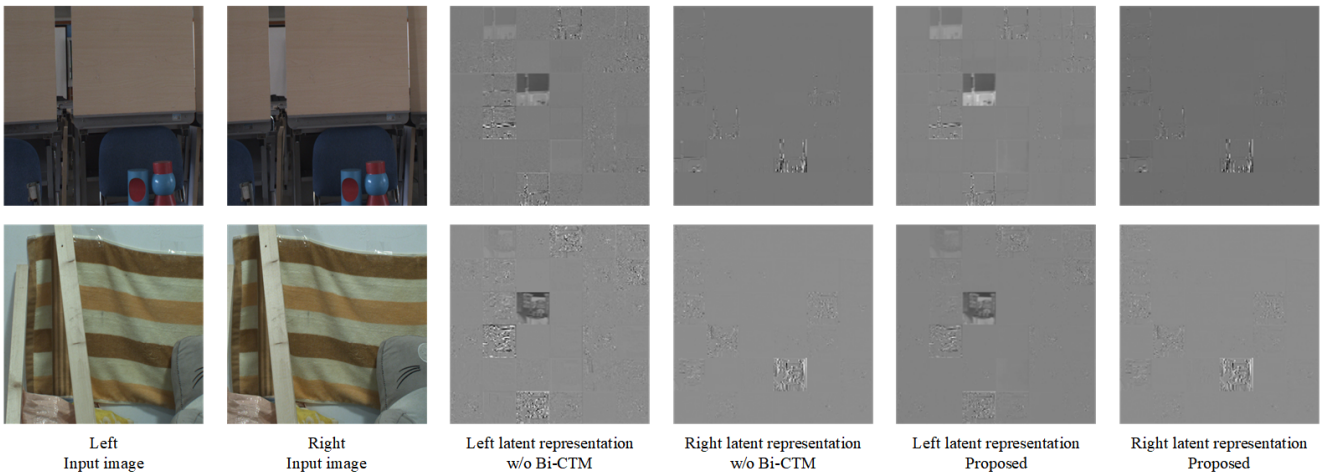

Figure 8. Visualization of the left and right latent representation of the proposed BCSIC-Net and BCSIC-Net without Bi-CTM. We use two stereo images from the InStereo2K test set, and assemble the visualization of the first 36 channels of their latent representation to display.

Table 4. Complexity comparison on the InStereo2K dataset.

Methods	FLOPs	Params	Enc-time	Dec-time
DSIC [24]	766.4 G	91.5 M	325.42 ms	246.31 ms
HESIC [13]	191.1 G	50.6 M	182.70 ms	8863.36 ms
Proposed	547.2 G	28.6 M	526.60 ms	12014.42 ms

erated in the two cases is similar. This is because the proposed Bi-CTM can jointly promote the compactness of both left and right latent representation by using the inter-view context.

2) Effectiveness of the Bi-directional Conditional Entropy Model. The proposed Bi-CEM aims for accurate probability estimation based on the inter-view correspondence. To verify its effectiveness, modifications similar to those in the ablation study of Bi-CTM are applied. That is, $p_{\hat{y}_L}(\hat{y}_L)$ is estimated without leveraging any right-view information. The results are also shown in Figure. 7 and Table 3, denoted as "w/o Bi-CEM". It can be seen that, if the bi-directional dependency is removed, the BD-PSNR performance decreases from 1.778 dB to 1.187 dB and the BD-rate performance decreases from 45.745% to 32.837%. This is because that the left-view probability estimation can no longer utilize the inter-view prior, resulting in a decrease in coding performance.

4.4. Complexity Analysis

Table 4 illustrates the complexity comparison between the proposed BCSIC-Net and the other two end-to-end stereo image compression methods. Compared with DSIC [24] and HESIC [13], model size of the proposed BCSIC-Net is reduced by 68% and 43% respectively, which shows that the proposed BCSIC-Net can provide competitive performance with smaller model size. As for the computation complexity, the FLOPs of the proposed BCSIC-Net is 5x higher than HESIC [13], but still lower than DSIC [24]. In addition, three stereo image compression methods are all tested on GTX 1080Ti to record the coding time. As shown in Table 4, the average coding time of the proposed BCSIC-Net is higher than that of DSIC [24] and HESIC [13].

5. Conclusion

This paper proposes a novel end-to-end stereo image compression network based on bi-directional coding (BCSIC-Net). In particular, a bi-directional contextual transform module is proposed to effectively reduce the inter-view redundancy by performing nonlinear transform conditioned on the inter-view context. In addition, a bi-directional conditional entropy model is developed to improve the accuracy of probability estimation for entropy coding by leveraging the inter-view correspondence as a prior. Experimental results demonstrate that the proposed BCSIC-Net achieves promising compression performance and is superior to the state-of-the-art methods.

Note that the decoding time of the proposed BCSIC-Net is longer than that of other state-of-the-art methods. This is mainly because the Bi-CEM with autoregressive context needs to decode all the pixels sequentially. According to the checkboard context model in [16], this issue can be alleviated by designing a parallel conditional entropy model for stereo image compression. We will work on this in the future.

Acknowledgements

This work was supported in part by the National Key R&D Program of China (No.2018YFE0203900), National Natural Science Foundation of China (No.61931014, 62125110), and Natural Science Foundation of Tianjin (No.18JCJQJC45800).

References

[1] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. Density modeling of images using a generalized normalization transformation. In *Proceedings of the International Conference on Learning Representations*, pages 1–14, 2016.

[2] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *Proceedings of the International Conference on Learning Representations*, pages 1–12, 2017.

[3] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *Proceedings of the International Conference on Learning Representations*, pages 1–13, 2018.

[4] Johannes Ballé, Philip A. Chou, David Minnen, Saurabh Singh, Nick Johnston, Eirikur Agustsson, Sung Jin Hwang, and George Toderici. Nonlinear transform coding. *IEEE Journal of Selected Topics in Signal Processing*, 15(2):339–353, 2020.

[5] Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. CompressAI: A PyTorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*, 2020.

[6] Fabrice Bellard. BPG image format. https://bellard.org/bpg/, 1, 2015.

[7] Gisle Bjontegaard. Calculation of average PSNR differences between RD-curves. *VCEG-M33*, 2001.

[8] Nikolaos V Boulgouris and Michael G Strintzis. A family of wavelet-based stereo image coders. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(10):898–903, 2002.

[9] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (VVC) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021.

[10] Tong Chen, Haojie Liu, Zhan Ma, Qiu Shen, Xun Cao, and Yao Wang. End-to-end learnt image compression via non-local attention optimization and improved context modeling. *IEEE Transactions on Image Processing*, 30:3179–3191, 2021.

[11] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. DSGN: Deep stereo geometry network for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12536–12545, 2020.

[12] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7939–7948, 2020.

[13] Xin Deng, Wenzhe Yang, Ren Yang, Mai Xu, Enpeng Liu, Qianhan Feng, and Radu Timofte. Deep homography for efficient stereo image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1492–1501, 2021.

[14] J.N. Ellinas and M.S. Sangriotis. Stereo image compression using wavelet coefficients morphology. *Image and Vision Computing*, 22(4):281–290, 2004.

[15] Tamas Frajka and Kenneth Zeger. Residual image coding for stereo image compression. *Optical Engineering*, 42(1):182–189, 2003.

[16] Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin. Checkerboard context model for efficient learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14771–14780, 2021.

[17] Yueyu Hu, Wenhan Yang, and Jiaying Liu. Coarse-to-fine hyper-prior modeling for learned image compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11013–11020, 2020.

[18] Mounir Kaaniche, Amel Benazza-Benyahia, Béatrice Pesquet-Popescu, and Jean-Christophe Pesquet. Vector lifting schemes for stereo image coding. *IEEE Transactions on Image Processing*, 18(11):2463–2475, 2009.

[19] Aysha Kadaikar, Gabriel Dauphin, and Anissa Mokraoui. Joint disparity and variable size-block optimization algorithm for stereoscopic image compression. *Signal Processing: Image Communication*, 61:1–8, 2018.

[20] Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. In *Proceedings of the International Conference on Learning Representations*, pages 1–12, 2019.

[21] Jianjun Lei, Zhe Zhang, Xiaoting Fan, Bolan Yang, Xinxin Li, Ying Chen, and Qingming Huang. Deep stereoscopic image super-resolution via interaction module. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8):3051–3061, 2021.

[22] Mu Li, Wangmeng Zuo, Shuhang Gu, Debin Zhao, and David Zhang. Learning convolutional networks for content-weighted image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3214–3223, 2018.

[23] Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, and Chunxia Xiao. ARShadowGAN: Shadow generative adversarial network for augmented reality in single light scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8139–8148, 2020.

[24] Jerry Liu, Shenlong Wang, and Raquel Urtasun. DSIC: Deep stereo image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3136–3145, 2019.

[25] Haichuan Ma, Dong Liu, Ning Yan, Houqiang Li, and Feng Wu. End-to-end optimized versatile image compression with wavelet-like transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, Doi:10.1109/TPAMI.2020.3026003.

[26] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Conditional probability models for deep image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4394–4402, 2018.

[27] David Minnen, Johannes Ballé, and George Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 10794–10803, 2018.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, and Luca Antiga. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 8026–8037, 2019.

[29] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient Attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3531–3539, 2021.

[30] Athanassios Skodras, Charilaos Christopoulos, and Touradj Ebrahimi. The JPEG 2000 still image compression standard. *IEEE Signal Processing Magazine*, 18(5):36–58, 2001.

[31] Xibin Song, Peng Wang, Dingfu Zhou, Rui Zhu, Chenye Guan, Yuchao Dai, Hao Su, Hongdong Li, and Ruigang Yang. ApolloCar3D: A large 3D car instance understanding benchmark for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5452–5462, 2019.

[32] Gary J. Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12):1649–1668, 2012.

[33] G.K. Wallace. The JPEG still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1):18–34, 1992.