Disparity-based Stereo Image Compression with Aligned Cross-View Priors

Yongqi Zhai†
zhaiyongqi@stu.pku.edu.cn
School of Electronic and Computer Engineering, Peking University
Shenzhen, China

Luyang Tang†
tly926@stu.pku.edu.cn
School of Electronic and Computer Engineering, Peking University
Shenzhen, China

Yi Ma
mayi@pku.edu.cn
School of Electronic and Computer Engineering, Peking University
Shenzhen, China

Rui Peng
ruipeng@stu.pku.edu.cn
School of Electronic and Computer Engineering, Peking University
Shenzhen, China

Ronggang Wang†
rgwang@pkusz.edu.cn
School of Electronic and Computer Engineering, Peking University
Shenzhen, China

ABSTRACT
With the wide application of stereo images in various fields, the research on stereo image compression (SIC) attracts extensive attention from academia and industry. The core of SIC is to fully explore the mutual information between the left and right images and reduce redundancy between views as much as possible. In this paper, we propose DispSIC, an end-to-end trainable deep neural network, in which we jointly train a stereo matching model to assist in the image compression task. Based on the stereo matching results (i.e., disparity), the right image can be easily warped to the left view, and only the residuals between the left and right views are encoded for the left image. A three-branch auto-encoder architecture is adopted in DispSIC, which encodes the right image, the disparity map and the residuals respectively. During training, the whole network can learn how to adaptively allocate bitrates to these three parts, achieving better rate-distortion performance at the cost of a lower disparity map bitrate. Moreover, we propose a conditional entropy model with aligned cross-view priors for SIC, which takes the warped latents of the right image as priors to improve the accuracy of the probability estimation for the left image. Experimental results demonstrate that our proposed method achieves superior performance compared to other existing SIC methods on the KITTI and iInStereo2K datasets both quantitatively and qualitatively.

CCS CONCEPTS
• Computing methodologies → Image compression.

†Both authors contributed equally to this research.
†Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:
Yongqi Zhai, Luyang Tang, Yi Ma, Rui Peng, and Ronggang Wang. 2022. Disparity-based Stereo Image Compression with Aligned Cross-View Priors. In Proceedings of the 50th ACM International Conference on Multimedia (MM ’22), October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3548136

1 INTRODUCTION
Recently, stereo images have been widely used in various fields, such as 3D movies, virtual reality, autonomous driving and so on. These large numbers of high-quality stereo images pose new challenges for data transmission and storage, which require efficient image compression methods to reduce the cost.

Image compression is a fundamental and vital research topic in the field of multimedia. After decades of research, a lot of traditional image compression standards have been proposed, such as JPEG [37], JPEG2000 [14], BPG [10] and VVC-intra [53]. Typically, they follow the pipeline of three hand-crafted modules: transformation, quantization and entropy coding. Recently, benefiting from the end-to-end optimization, deep neural network (DNN) based image compression methods have achieved promising results [3, 4, 12, 21, 23, 32, 45, 47, 58, 65].

A stereo image pair consists of left and right images, which are captured by a stereo camera from both views at the same time. For stereo images, a simple idea is to compress each image separately using a single image compression method. However, due to the large overlapping fields of view between the left and right cameras, quite a lot of similar information inherently exists between stereo images. This idea ignores the high correlation between views and wastes a lot of bitrates, which is rather unwise. Therefore, it is necessary to design a compression method that can make full use of the inner relationship and minimize the redundancy between views as much as possible.

Stereo image compression (SIC) compresses the left and right images jointly, aiming to achieve high compression ratios for both two images. At present, there have been many studies focusing on SIC. In many traditional SIC methods [36, 44], the disparity map is often compressed additionally as important side information to reduce
the bitrates. Given a reference view, they use the disparity compensated prediction to encode other views. However, traditional methods minimize the rate-distortion loss based on hand-crafted features, which limits their compression efficiency. Recently, deep learning-based SIC methods have achieved better performance than traditional ones. Liu et al. [35] proposed the first deep stereo image compression network, namely DSIC, which densely warps the features from the left image to the right at all levels in the encoder and decoder to exploit the content redundancy between the stereo pair. When training and testing, DSIC requires to construct cost volume will make the residuals in HESIC very large. Compared with the (HESIC), which reduces the computation complexity and achieves better compression performance. In HESIC, the left image is spatially transformed by a 3 × 3 homography matrix and the residuals between the right and the transformed left image are encoded. However, the homography matrix is used to describe the relationship between two images of the same planar surface in space. Most stereo image pairs have difficulty meeting this condition, which will make the residuals in HESIC very large. Compared with the homography matrix, the disparity map establishes a more accurate pixel-to-pixel correspondence between stereo images, which can significantly reduce the residual information. Moreover, as a simple 1D data flow, the disparity map is easily compressed.

In this paper, we propose a disparity-based deep neural network for stereo image compression, namely DispSIC, which combines the disparity compensated prediction in traditional methods with a deep learning framework. As Fig. 1 shows, we replace the homography matrix in HESIC [16] with the disparity map, which is output by a stereo matching model. Unlike the traditional SIC methods using disparity compensated prediction, we train the stereo matching model and the compression model jointly with the rate-distortion loss. By end-to-end optimization, the stereo matching model learns how to provide more accurate disparity compensated prediction, and the compression model learns how to trade off the transmission cost and gain of the disparity map. The two models work toward the same objective i.e. reconstructing higher quality images with fewer bitrates.

In conclusion, the main contributions of this paper are as follows:

- We propose a novel disparity-based deep neural network for stereo image compression, namely DispSIC, which uses the disparity map to explicitly represent the relationship between the left and right images.
- We design a conditional entropy model with aligned cross-view priors, which can model the probability of the left image more accurately.
- The experimental results show that our proposed method significantly outperforms the state-of-art deep stereo image compression methods, and saves around 25% bitrates compared to the latest method HESIC [16] with similar image quality.

2 RELATED WORK

2.1 Single Image Compression

Existing traditional image compression standards include JPEG [57], JPEG2000 [14], BPG [10] and VVC-intra [53]. The pipeline of these methods typically consists of transformation, quantization, and entropy coding. Although these traditional methods can achieve high compression efficiency, they heavily rely on prior knowledge to design hand-crafted modules. In recent years, the DNN-based image compression methods [3, 4, 12, 21, 23, 32, 45, 47, 58, 65] have achieved great success with impressive performance. Unlike traditional methods, they jointly optimize all modules in an end-to-end manner. For the network architecture, some early works used the recurrent neural networks (RNNs) to achieve progressive compression [52, 56]. Then, most works used the convolutional neural networks (CNNs) to build an auto-encoder style network. Most recently, there have been some works employing the invertible neural networks (INNs) [58] and transformer [47, 65] architecture. Besides, some works [1, 3, 55] were proposed to solve the problem of non-differential quantization and rate estimation. Other works such as generalized divisive normalization (GDN) [3], attention mechanism [12, 34], residual blocks [12], adaptive feature extraction models [38], adversarial training [2, 43, 48], and importance map [42, 63] were focused on improving the image compression performance. Meanwhile, variable-rate model [13, 15], scalable compression [25, 41] were introduced to meet the practical application requirements.

2.2 Stereo Matching

Stereo matching is the process of finding pixels corresponding to the same 3D point in the scene, consisting of four steps: matching cost computation, cost aggregation, disparity optimization and post-processing [50]. As we know, the core of SIC is to explore the inner correlation between the left and right images, in which stereo matching techniques play an important role.

Traditional stereo matching methods can be roughly divided into three categories: local methods [8, 49], global methods [30, 31] and semi-global methods [22]. Local methods compute the matching cost on a block-by-block or pixel-by-pixel basis, which run fast but have low-quality results. Global methods first obtain an initial
Disparity-based Stereo Image Compression with Aligned Cross-View Priors

2.3 Stereo Image Compression

Traditional SIC methods [7, 17, 18, 26, 29, 39, 44] usually compress stereo images in the way of video compression, which regards the inter-view similarity between the stereo camera views as the temporal similarity between successive frames in the video. Given a reference view, many works [36, 44] use the disparity compensated prediction to encode each view, which is similar to the motion-compensated prediction in the single-view video. These methods rely on hand-crafted features and design modules individually, which greatly limits their compression performance.

Recently, several DNN-based methods for SIC have been proposed, which outperform the traditional methods. Liu [35] proposed the first DNN-based stereo image compression framework (DSIC), which uses parametric skip functions to feed the disparity-warped features at all levels from the left encoder/decoder to the right one. Due to the dense warp scheme, it has high computation complexity. After that, Deng et al. [16] use a homography transformation to replace the dense warp, which reduces the computation complexity and achieves state-of-art performance. In HESIC [16], a 3x3 homography matrix is used to spatially transform the left image to the right view. And only the residuals between the original right image and the transformed left image are encoded to reconstruct the right image. However, only when two images are of the same planar surface in space, the homography matrix can provide an accurate match. So in most scenarios, the homography transformation will lead to large residuals in HESIC [16].

3 METHODS

3.1 Framework

Fig. 2 shows the overall framework of our proposed DispSIC network. In general, the whole network adopts a three-branch structure, which is used to encode the right image, the disparity map and the residuals of the left image respectively.

Firstly, taking the original left and right images (denoted as \(x_l\) and \(x_r\)) as input, we use a stereo matching model to estimate the disparity map (denoted as \(d_l\)). Then, \(x_l\) and \(d_l\) are compressed via two independent auto-encoders [45]. Based on the decoded disparity map \(d_l\), the decoded right image \(\hat{x}_r\) can be easily warped...
to the left view, obtaining the predicted left image \(\hat{x}_{r \rightarrow l}\). After that, we concatenate the original left image \(x_l\) and the predicted \(\hat{x}_{r \rightarrow l}\) as the input to the third auto-encoder [45], which learns to compress the residuals. Taking into account the correlation between \(x_l\) and \(x_r\), we take the warped quantized latent of \(x_r\) as priors for the entropy model of the latent of \(x_l\). At the decoder side, the decoded residuals are concatenated with the predicted \(\hat{x}_{r \rightarrow l}\) to reconstruct the final left image \(\hat{x}_l\).

### 3.2 Disparity-based Prediction

To fully explore the inner relationship between stereo images, we need to match the images or features first. DSIC [35] constructs cost volume at each feature level to match the features between stereo images, which works but has high computation complexity. HESIC [16] uses a \(3 \times 3\) homography matrix to roughly match the stereo images, which is simple but inaccurate. As Fig. 3 shows, the image generated by the homography transformation has many problems, such as border pixels missing and misaligned textures with the original image, which greatly increase the residual information.

In contrast, the disparity refers to the difference in coordinates of similar features within two stereo images, which can reflect the pixel-level correlation between the left and right images. As a 1D data flow, it is lightweight and can be easily compressed. Fig. 3 shows that the image generated by disparity-based warping is of better quality and has less residual information. Therefore, we propose to use the disparity map to assist with stereo image compression.

For convenience, we adopt an existing stereo matching model [59]. Taking the left image \(x_l\) and the right image \(x_r\) as input, it can generate the disparity map \(d_l\) aligned with \(x_l\). On the decoder side, based on the decoded disparity map \(d_l\), following [24] we can obtain the left predicted image \(\hat{x}_{r \rightarrow l}\) by using a bilinear sampler to sample the decoded right image \(\hat{x}_r\) with backward mapping. The whole process is fully differentiable and can integrate into our end-to-end trainable compression network.

### 3.3 Aligned Cross-View Priors

Given the latent \(\hat{y}_r\), the entropy model \(p_y\) is used to fit the marginal distribution \(m(\hat{y}_l)\). The smallest average code length (bitrates) is decided by the matching degree of \(p_y\) and \(m(\hat{y}_r)\), given by the Shannon cross entropy between the two distributions:

\[
R = \mathbb{E}_{\hat{y}_l \sim m}[-\log_2 p_y(\hat{y}_l)].
\]

For the quantized hyper latent \(\hat{z}_l, \hat{z}_r\), we use a non-parametric, fully factorized density model [3]. For the right quantized latent \(\hat{y}_r\) and disparity map quantized latent \(\hat{y}_d\), we use an autoregressive context model with a mean and scale hyperprior [45], which can be formulated as:

\[
\begin{align*}
    p_{\hat{y}_l|\hat{z}_l} (\hat{y}_l|\hat{z}_l) & \sim \mathcal{N}(\mu_l, \sigma_l^2), \\
    p_{\hat{y}_d|\hat{z}_d} (\hat{y}_d|\hat{z}_d) & \sim \mathcal{N}(\mu_d, \sigma_d^2),
\end{align*}
\]

where \(\mu, \sigma\) are the estimated mean and scale parameters.

Due to the large overlapping fields of view between the left and right cameras, strong correlation exists between the left quantized latent \(\hat{y}_l\) and right quantized \(\hat{y}_r\). We visualize the features of the left and right images in Figure 4. It is obvious that \(\hat{y}_l\) and \(\hat{y}_r\) have structural similarities, so the mutual information \(I(\hat{y}_l, \hat{y}_r)\) is positive. We denotes the Shannon entropy of \(\hat{y}_l\) as \(H(\hat{y}_l)\), and then the
where $R$ denotes the bitrates, $D$ denotes the distortion and $\lambda$ controls the rate-distortion trade-off.

To stabilize the training process, we use the pre-trained stereo matching model [59] as the initialization of disparity estimation. In addition, we introduce an auxiliary loss $D(x_i, \hat{x}_{i-1})$ at the early stage of training to instruct the whole network to utilize the disparity compensated prediction. The initial loss function $L_{\text{init}}$ is defined as follows:

$$L_{\text{init}} = R_l + R_r + R_d + \lambda (D(x_i, \hat{x}_i) + D(x_r, \hat{x}_r) + \beta D(x_i, \hat{x}_{i-1})),$$

where the hyperparameter $\beta$ is set to 0.2. When the training loss $L_{\text{init}}$ is stable, the auxiliary loss $D(x_i, \hat{x}_{i-1})$ is removed and $\beta$ is set to 0. The total rate-distortion loss $L_{\text{final}}$ can be formulated as:

$$L_{\text{final}} = R_l + R_r + R_d + \lambda (D(x_i, \hat{x}_i) + D(x_r, \hat{x}_r)).$$

### 4 EXPERIMENTS

#### 4.1 Experimental Setup

**Datasets.** We evaluate the compression performance of our proposed DispSIC on two public stereo image datasets:

1. KITTI [19]: A dataset of outdoor scenes. The stereo images in it are with far views. We use 1,950, 50 and 50 stereo image pairs which are randomly selected by HESIC [16] for training, validation and testing respectively.

2. InStereo2K [5]: A dataset of indoor scenes. It consists of 2050 stereo image pairs. The stereo images in it are with close-views. Likewise, 1,950 pairs are used for training and 50 pairs for validation. The remaining 50 pairs are used for testing.

During training, the images are randomly cropped to 256 × 512 patches. By using these two datasets of different scenarios, we can evaluate our method comprehensively.

**Implementation details.** We implement our framework on the CompressAI PyTorch library [6]. We optimize our network using Adam optimizer [28] with a batch size of 4. We first train our model on KITTI, and then the model is fine-tuned to InStereo2K. All models are trained on a single V100 GPU for 1600 epochs. The learning rate is $1 \times 10^{-4}$ for 1200 epochs and reduced to $1 \times 10^{-5}$ for the last iterations. We train our network with the loss $L_{\text{init}}$ in the first 500 epochs, and then with the loss $L_{\text{final}}$ in the remaining epochs. We use mean square error (MSE) as our quality metric, and train 5 models for both datasets. The $\lambda$ is chosen from set [0.001, 0.002, 0.005, 0.01, 0.02].

**Evaluation.** To evaluate the rate-distortion performance, we compare our model DispSIC with the learned stereo image compression methods [16, 35], the SOTA learned single image compression methods [12, 23], and the traditional image and video compression codecs [10, 53, 54] on the KITTI and InStereo2K dataset. It is worth mentioning that when comparing with HEVC [54], stereo images are fed into the encoder as video sequences. The rate is measured by bits per pixel (bpp) and the quality is measured by peak signal-to-noise ratio (PSNR) and multi-scale structural similarity (MS-SSIM). Note the bpp is the average of both images. To compare the coding efficiency of different methods, we draw the rate-distortion (RD) curves. In addition, we report the Bjontegaard delta PSNR (BD-PSNR) [9] and BD-rate results to better compare methods.

---

**Figure 5:** Our conditional entropy model used to encode the quantized latent $\hat{y}_i$. ENC* and DEC* represent the hyper-prior encoder and decoder. AE and AD are the arithmetic encoder and arithmetic decoder.
Figure 6: Rate-distortion performance comparison of different methods on the KITTI and InStereo2K datasets.

Table 1: BD-PSNR and BD-rate comparisons, with the best results in red.

| KITTI dataset                        | BD-PSNR(dB) | BD-rate(%) |
|--------------------------------------|-------------|------------|
| Methods                              |             |            |
| BPG                                  | -0.533      | 22.624     |
| HEVC/H.265                           | -0.412      | 15.665     |
| Hu (AAAI’20)                         | 0.004       | 5.510      |
| HESIC (CVPR’21)                      | 0.357       | -13.296    |
| Cheng (CVPR’20)                      | 0.234       | -7.723     |
| VVC-intra-444                        | 0.691       | -23.113    |
| HESIC+ (CVPR’21)                     | 0.862       | -27.069    |
| DispSIC                              | 2.308       | -51.719    |

| InStereo2K dataset                   | BD-PSNR(dB) | BD-rate(%) |
|--------------------------------------|-------------|------------|
| Methods                              |             |            |
| BPG                                  | -0.478      | 20.683     |
| HEVC/H.265                           | -0.043      | 0.677      |
| Hu (AAAI’20)                         | 0.420       | -15.108    |
| HESIC (CVPR’21)                      | 0.625       | -22.19     |
| Cheng (CVPR’20)                      | 0.681       | -23.666    |
| VVC-intra-444                        | 0.903       | -30.171    |
| HESIC+ (CVPR’21)                     | 0.838       | -28.920    |
| DispSIC                              | 1.847       | -53.544    |

higher value of BD-PSNR and the lower value of BD-rate indicate better image compression performance.

Table 2: Computational complexity comparison.

| Method   | FLOPs   | Params |
|----------|---------|--------|
| DSIC     | 178.4G  | 91.5M  |
| HESIC+   | 48.6G   | 50.6M  |
| DispSIC  | 43.5G   | 15.9M  |

Table 3: The specific configuration of ablation study.

| Disparity | Priors | Aligned Priors | PRN |
|-----------|--------|----------------|-----|
| Case 1    |        |                |     |
| Case 2    | ✓      |                |     |
| Case 3    | ✓      | ✓              |     |
| Case 4    | ✓      | ✓              | ✓   |
| DispSIC   | ✓      | ✓              | ✓   |

4.2 Compression Results of Proposed Methods

Quantitative results. Figure 6 shows the rate-distortion performance of different methods on the KITTI and InStereo2K datasets. The data points on the RD curves are collected from the official GitHub page of HESIC [16]. For the MS-SSIM curve, in addition to the original scale, we also draw the log-scale, defined as $-10 \log_{10}(1-\text{MS-SSIM})$. Compared with the state-of-art (SOTA) stereo image
Figure 7: Qualitative comparison. It is worth mentioning that these images are not cherry-picked, they are the same as images in HESIC [16].

compression method HESIC+ [16], the SOTA learned image compression method Cheng [12], and the SOTA traditional image compression codec VVC-intra [53], our method achieves the best performance on both datasets, which demonstrates the effectiveness and generality of our method. It is worth mentioning that our method achieves greater gain on the KITTI dataset at high bitrates. This demonstrates that our disparity-based method is more effective in scenes with complex structures. Table 1 shows the BD-BR and BD-PSNR results with DSIC [35] as the baseline, we can see that our method achieves the highest BD-PSNR and the lowest BD-rate. 

Qualitative results. In Fig. 7, we provide some reconstructed images of different methods on the InStereo2K dataset. For a fair comparison, all images are compressed to similar bit rates. As we can see, our model achieves higher PSNR of reconstructed images at lower bitrates. Meanwhile, it can be seen that our method can generate more structural details than all other methods, such as the stripes on the tiger and the lines on the basketball.

Computational complexity. In Table 2, we compare the FLOPs and parameters of DispSIC with HESIC+ and DSIC. Specifically, the resolution of the input image is $256 \times 256$. It can be seen that the Flops of DispSIC are slightly smaller than HESIC+ and much smaller than DSIC, and the number of parameters is much smaller than both HESIC+ and DSIC. The dense warp scheme of DSIC and the autoregressive model of HESIC+ for generating the H matrix have high computational complexity. Instead, we use a lightweight stereo matching model to explore the mutual information between stereo images, which greatly reduces the computational complexity. All in all, our model DispSIC has better performance and less complexity.
Table 4: Rate allocation of DispSIC on the KITTI dataset.

| Dataset | BPP  | PSNR | BPP_R | PSNR_R | BPP_L | PSNR_L | BPP_D | PSNR_R-L |
|---------|------|------|-------|--------|-------|--------|-------|----------|
| KITTI   | 0.177| 26.45| 0.222 | 27.30  | 0.118 | 25.60  | 0.014 | 21.170   |
|         | 0.320| 28.60| 0.380 | 29.35  | 0.242 | 27.86  | 0.018 | 21.493   |
|         | 0.488| 30.19| 0.560 | 30.97  | 0.395 | 29.41  | 0.022 | 21.552   |
|         | 0.701| 31.72| 0.775 | 32.55  | 0.598 | 30.90  | 0.028 | 21.678   |

4.3 Analysis of Rate Allocation

In Table 4, we count the rate allocation of our method at different bitrates in the RD-curves on the KITTI dataset. Note the calculation of $BPP_L$, $BPP_R$, and $BPP_D$ is the bitstream size divided by the number of pixels in a single image, and $PSNR_{R→L}$ denotes the PSNR of the left disparity-based prediction.

As we can see, our network assigns the most bitrates ($BPP_R$) to the right image. As to the left image, we compress the left residuals and the disparity map. The bitrates for the left residuals ($BPP_L$) are less. Moreover, the disparity map is assigned the least bitrates ($BPP_D$). Nonetheless, it can provide a respectable disparity compensated prediction, which greatly reduces the left residuals. In conclusion, by end-to-end optimization, our network can trade off the transmission cost and gain of the disparity map and adaptively allocate bitrates to the three parts to achieve the optimal compression performance.

4.4 Ablation Study

To analyze the contribution of each module, we implement ablation experiments on the KITTI and InStereo2K datasets as Figure 8 shows. Table 3 shows the specific configuration of ablation study.

**Disparity-based Prediction.** In case 1, we remove the disparity-based prediction and the priors. That means the stereo images are fed to the encoder separately, which can be considered as single image compression. As shown in Fig. 8, removing the disparity-based prediction causes significant performance degradation (case 1 vs. case 2), which demonstrates that the disparity-based prediction is the core module of our framework.

**Aligned Cross-View Priors.** To exploit the effectiveness of the aligned cross-view priors, we implement three experiments. In case 2, we simply remove the priors and use two independent entropy models for left and right images. In case 3, cross-view priors are sent to the left image entropy model, but lack alignment operation. In case 4, we align the cross-view priors with the left features. As we can see, the RD performance has a large drop if priors are disabled (case 2 vs. case 3). When enabling aligned cross-view priors (case 3 vs. case 4), the RD performance can be further improved, especially in scenes more complex. These two comparative experiments demonstrate the effectiveness of our aligned cross-view priors.

**Prior Refinement Net.** Since the warping operation may introduce some spatial discontinuity, we design a prior refinement net to further improve the quality of the aligned cross-view priors. As shown in Fig. 8, the RD curve of case 4 is lower than the original model DispSIC. This result shows that the prior refinement net is helpful to improve the quality of the priors.

5 CONCLUSION

In this paper, we propose a disparity-based stereo image compression network, which outperforms other existing SIC methods. Based on the disparity map, the right image is warped to the left view to get the residual image. Then the right image, the disparity map and the residual image are encoded into bitstream. Moreover, we propose a conditional entropy model with aligned cross-view priors, which provides more accurate probability estimation. Experiments on the KITTI and InStereo2K datasets show that our algorithm outperforms state-of-art learned stereo image compression methods.
ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China U21B2012, 62072013 and 61902008, Shenzhen Cultivation of Excellent Scientific and Technological Innovation Talents BRC2020071414435057, Shenzhen Research Projects of JCYJ2018 0503182128089 and 20180608092149290.

REFERENCES

[1] Eirikur Agustsson, Fabian Mentzer, Michael Tscharnien, Lukas Cavigelli, Radu Timofte, Luca Benni, and Luc Van Gool. 2017. Soft-to-hard vector quantization for end-to-end learning compressible representations. arXiv preprint arXiv:1704.08648 (2017).

[2] Eirikur Agustsson, Michael Tscharnien, Fabian Mentzer, Radu Timofte, and Luc Van Gool. 2019. Generative adversarial networks for extreme learned image compression. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 221–231.

[3] Johannes Ballé, Valero Laparra, and Zdeněk P. Simoncelli. 2017. End-to-end Opti- mized Image Compression. (2017).

[4] Johannes Ballé, David Minnen, Sohrab Singh, Sung Jin Hwang, and Nick Johnson. 2018. Variational image compression with a scale hyperprior. arXiv preprint arXiv:1802.01436 (2018).

[5] Wei Bao, Wei Wang, Tuhua Xu, Yulan Guo, Siyu Hong, and Xiaohu Zhang. 2020. InStereo2K: A large real dataset for stereo matching in indoor scenes. Science China Information Sciences 63, 11 (2020), 1–11.

[6] Jean Béjart, Fabien Racapé, Simon Fillman, and Akshay Pushparaj. 2020. Compressia: a pytorch library and evaluation platform for end-to-end compression research. arXiv preprint arXiv:2011.03629 (2020).

[7] I Berriz, Moumine Kaaniche, Saadi Boudjed, and Azeddine Beghdadi. 2018. Sparse optimization of non separable vector lifting scheme for stereo image coding. Journal of Visual Communication and Image Representation 57 (2018), 283–293.

[8] Stan Birchfield and Carlo Tomasi. 1999. Depth discontinuities by pixel-to-pixel stereo. International Journal of Computer Vision 35, 3 (1999), 269–293.

[9] Gile Bjontegaard. 2001. Calculation of average PSNR differences between RD-curves. VCEG-M33 (2001).

[10] BPG. 2019. BPG Image format. https://bellard.org/bpg/.

[11] Jia-Ren Chang and Yong-Sheng Chen. 2018. Pyramid stereo matching network. arXiv preprint arXiv:1802.01436 (2017).

[12] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. 2020. Learned wavelet coefficients morphology. IEEE Transactions on Pattern Analysis and Machine Intelligence 42, 10 (2020), 1–8.

[13] Alex Kendall, Havik Martirosyan, Sammito Daguapeta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. 2017. End-to-end learning of geometry and context for deep stereo regression. In Proceedings of the IEEE international conference on computer vision. 66–75.

[14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

[15] Masaki Kitahara, Hideaki Kimata, Shinya Shinm, Kazuno Kamikura, Yoshuki Yoshima, Kenji Yamamoto, Tomohito Yendo, Toshihiko Fujii, and Masayuki Tanimoto. 2006. Multi-view video coding using view interpolation and reference picture selection. In 2006 IEEE International Conference on Multimedia and Expo. IEEE, 97–100.

[16] Andreas Klaus, Mario Sormann, and Konrad Karner. 2006. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In 18th International Conference on Pattern Recognition (ICPR’06). Vol. 3. IEEE, 15–18.

[17] Vladimír Kolmogorov and Ramin Zabih. 2001. Computing visual correspondence with occlusions using graph cuts. In Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, Vol. 2. 508–515.

[18] Jooyoung Lee, Seunghyun Cho, and Seungkwan Beak. 2019. Context-adaptive Entropy Model for End-to-End Optimized Image Compression. (2019).

[19] Zhengfa Liang, Yilu Feng, Yulan Guo, Hengzhao Liu, Wei Chen, Limbo Qiao, Li Zhou, and Jianfeng Zhang. 2018. Learning for disparity estimation through feature constancy. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2011–2020.

[20] Haojie Liu, Tong Chen, Peiyao Guo, Qui Shen, Xun Cao, Yao Wang, and Zhuan Ma. 2019. Non-local attention optimized deep image compression. arXiv preprint arXiv:1904.09759 (2019).

[21] Jerry Liu, Shemlong Wang, and Raquel Urtasun. 2019. DSC: Deep stereo image compression. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 3136–3145.

[22] M Lukacs. 1986. Predictive coding of multi-viewpoint image sets. In ICASSP’86. IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 11. IEEE, 521–524.

[23] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. 2016. Efficient deep learning for stereo matching. In Proceedings of the IEEE conference on computer vision and pattern recognition. 5695–5703.

[24] Yi Ma, Yongzi Zhai, Jiayu Yang, Chunhui Yang, and Ronggang Wang. 2021. AFEPC: Adaptive Feature Extraction Modules for Learned Image Compression. In Proceedings of the 29th ACM International Conference on Multimedia. 5436–5444.

[25] Ewoud Martijn, Alexander Behrens, Jun Xun, and Anthony Vetro. 2006. View synthesis for multiview video compression. In Picture Coding Symposium, Vol. 37. 38–39.

[26] Nikolaus Mayer, Eddy Ilg, Philipp Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 4040–4048.

[27] Yuzin Mei, Li Li, Zhu Li, and Fan Li. 2021. Learning-Based Scalable Image Compression with Latent-Feature Reuse and Prediction. IEEE Transactions on Multimedia (2021).

[28] Jerry Liu, Shenlong Wang, and Raquel Urtasun. 2019. DSIC: Deep stereo image compression. arXiv preprint arXiv:1904.09759 (2019).

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

[30] Fabian Mentzer, George Toderici, Michael Tscharnien, and Eirikur Agustsson. 2020. High-fidelity generative image compression. arXiv preprint arXiv:2006.09965 (2020).

[31] Philipp Merkle, Karsten Muller, Aljoscha Smolic, and Thomas Wiegand. 2006. Efficient compression of multi-view video exploiting inter-view dependencies based on H.264 MPE-G4-FEC. In 2006 IEEE International Conference on Multimedia and Expo. IEEE, 1717–1720.

[32] David Minnen, Johannes Ballé, and George D Toderici. 2018. Joint autoregressive and hierarchical priors for learned image compression. In Advances in Neural Information Processing Systems. 10771–10780.

[33] Andrea Pöltner, Stéphane Lathuilière, Dan Xu, Mihai Marian Puscas, Elisa Ricci, and Nicu Sebe. 2019. Progressive fusion for unsupervised binocular depth estimation using cycled networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 42, 10 (2019), 2380–2395.
[47] Yichen Qian, Ming Lin, Xuexu Sun, Zhiyu Tan, and Rong Jin. 2022. Entroformer: A Transformer-based Entropy Model for Learned Image Compression. arXiv preprint arXiv:2202.05492 (2022).

[48] Oren Rippel and Lubomir Bourdev. 2017. Real-time adaptive image compression. In International Conference on Machine Learning. PMLR, 2922–2930.

[49] Thomas W Ryan, RT Gray, and Bobby R Hunt. 1980. Prediction of correlation errors in stereo-pair images. Optical Engineering 19, 3 (1980), 312–322.

[50] Daniel Scharstein and Richard Szeliski. 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International journal of computer vision 47, 1 (2002), 7–42.

[51] Xiao Song, Xu Zhao, Liangji Fang, Hanwen Hu, and Yuhou Yu. 2020. Edgestereo: An effective multi-task learning network for stereo matching and edge detection. International Journal of Computer Vision 128, 4 (2020), 910–930.

[52] Rige Su, Zhengxue Cheng, Heming Sun, and Jiroy Katto. 2020. Scalable Learned Image Compression With A Recurrent Neural Networks-Based Hyperprior. In 2020 IEEE International Conference on Image Processing (ICIP). IEEE, 3369–3373.

[53] G. J. Sullivan and J. R. Ohm. 2018. Versatile video coding Towards the next generation of video compression. Picture Coding Symposium (2018).

[54] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. 2012. Overview of the high efficiency video coding (HEVC) standard. IEEE Transactions on circuits and systems for video technology 22, 12 (2012), 1649–1668.

[55] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. 2017. Lossy image compression with compressive autoencoders. arXiv preprint arXiv:1703.00395 (2017).

[56] George Toderici, Sean M O’Malley, Sung Jin Hwang, Damien Vincent, David Minnen, Shumeet Baluja, Michele Covell, and Rahul Sukthankar. 2015. Variable rate image compression with recurrent neural networks. arXiv preprint arXiv:1511.06085 (2015).

[57] Gregory K. Wallace. 1992. The JPEG still picture compression standard. Communications of the ACM 35, 1 (1992), 86–94.

[58] Yueqi Xie, Ka Leong Cheng, and Qifeng Chen. 2021. Enhanced invertible encoding for learned image compression. In Proceedings of the 29th ACM International Conference on Multimedia. 162–170.

[59] Haofei Xu and Juyong Zhang. 2020. Aanet: Adaptive aggregation network for efficient stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1959–1968.

[60] Guorun Yang, Henghuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. 2018. Segstereo: Exploiting semantic information for disparity estimation. In Proceedings of the European conference on computer vision (ECCV). 636–651.

[61] Lidong Yu, Yucheng Wang, Yuwei Wu, and Yunde Jia. 2018. Deep stereo matching with explicit cost aggregation sub-architecture. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32.

[62] Jure Zbontar, Yann LeCun, et al. 2016. Stereo matching by training a convolutional neural network to compare image patches. J. Mach. Learn. Res. 17, 1 (2016), 2287–2318.

[63] Xi Zhang and Xiaolin Wu. 2021. Attention-guided Image Compression by Deep Reconstruction of Compressive Sensed Saliency Skeleton. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 13354–13364.

[64] Yiran Zhong, Yuchao Dai, and Hongdong Li. 2017. Self-supervised learning for stereo matching with self-improving ability. arXiv preprint arXiv:1709.08930 (2017).

[65] Yinzhao Zhu, Yang Yang, and Taco Cohen. 2022. Transformer-based Transform Coding. In International Conference on Learning Representations.