2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Cross View Capture for Distributed Image Compression with Decoder Side Information

Yankai Yin[1,2], Zhe Sun[3], Peiying Ruan[4], Feng Duan[1*], Ruidong Li[2*], Chi Zhu[5]

Abstract— **Image compression is increasingly important in applications like intelligent driving and smart surveillance systems. This study presents a novel cross view capture distributed image compression network (CVCDIC) to improve the compression quality by using decoder side information. The CVCDIC's decoder utilizes feature extraction networks to extract features from both the primary image and the side information. Furthermore, a multi-level cross view attention module is designed to capture interrelated details between images at multiple hierarchical levels. Finally, a spatial refinement module, constructed on the foundation of information distillation networks, is designed to further refine the quality of reconstructed images. The results show that CVCDIC can achieve an MS-SSIM of 0.978 at 0.15 bpp, surpassing DSIN (0.925), NDIC (0.956), and ATN (0.955) on the KITTI Stereo dataset.**

I. Introduction

The widespread adoption of intelligent robots and autonomous driving systems has led to an exponential increase in image data. High-resolution and high-frame-rate images from these systems pose significant challenges in terms of image storage, transmission, and processing. Image compression techniques aim to mitigate these challenges by reducing storage and bandwidth requirements while maintaining acceptable image reconstruction quality. Broadly, image compression algorithms can be classified into two categories: lossless and lossy compression [1]. Lossless compression methods, which retain all the original data, are mainly used in applications requiring high fidelity, such as medical imaging and fingerprint recognition. Classic lossless techniques include algorithms such as Portable Network Graphics (PNG), which primarily utilizes the Deflate algorithm, and Joint Photographic Experts Group-Lossless (JPEG-LS), which relies on localized prediction principles. Conversely, lossy compression is more commonly employed due to its ability to significantly increase compression ratios at the cost of acceptable quality degradation. Traditional lossy algorithms, including the JPEG, JPEG2000, High Efficiency Video Coding (HEVC), and Versatile Video Coding (VVC), generally divide the image into distinct blocks for easier processing. Learning-based image compression methods often treat the image as a whole. Learning-based approaches primarily rely on Variational Autoencoders (VAEs) for their underlying architecture [2]. In the learning-based image compression framework, the latent representation of the image is quantized and then entropy-encoded using a probability distribution learned by the entropy model. Learning-based techniques excel at adaptively capturing intricate data distributions in images, leading to more efficient encoding. Additionally, learning-based methods generally demonstrate superior precision during the image reconstruction phase.

In applications like autonomous driving and remote surveillance, stereo camera arrays are often used to capture a broader field of view collectively. Traditional stereo image compression algorithms usually require simultaneous access to correlated image pairs for joint encoding. This demands inter-device communication and data synchronization for joint encoding, which may not be practical for devices with limited resources. Distributed image compression offers the unique advantage of allowing independent encoding and collaborative decoding for correlated image streams [3]. In these scenarios, a more powerful decoding system can use side information from correlated images to improve image reconstruction quality or reduce the bit rate, thus optimizing overall compression efficiency.

In this study, we propose a novel framework that integrates features from both primary and correlated images in the decoding process, aiming to improve the rate distortion (RD) performance in image compression. The key contributions of this study are summarized as follows:

- A novel cross view capture distributed image compression network (CVCDIC) is proposed to effectively leverage both intra-image and inter-image information. The proposed method shows significant performance gains over existing methods in distributed image compression, as confirmed by tests on benchmark datasets like KITTI and Cityscapes.
- During decoding, multiple feature extraction networks are used to iteratively extract intrinsic features from both primary and side information images.
- A Multi-Level Cross View Attention Module (ML-CAM) is designed to fuse correlated information between image pairs at both epipolar line and patch levels.
- A Spatial Refinement Module (SRM), constructed on the information distillation structure, is designed to improve the quality of the reconstructed image.

This research was supported by the National Natural Science Foundation of China (Key Program) (No. 11932013), and the Tianjin Science and Technology Plan Project (No. 22PTZWHZ00040).

[1]Tianjin Key Laboratory of Interventional Brain-Computer Interface and Intelligent Rehabilitation, Nankai University, Tianjin 300350, China
[2]Institute of Natural Sciences, Kanazawa University, Ishikawa 9201164, Japan
[3]Faculty of Health Data Science and Graduate School of Medicine, Juntendo University, Chiba 2790013, Japan
[4]NVIDIA AI Technology Center, NVIDIA Japan, Tokyo 1070052, Japan
[5]Department of Systems Life Engineering, Maebashi Institute of Technology, Maebashi 3710816, Japan
*Corresponding authors: Feng Duan (duanf@nankai.edu.cn), Ruidong Li (liruidong@ieee.org)

II. Related Work

In this section, we detail discuss the image compression architecture and multi-view image compression method.

A. Image Compression Architecture

In recent years, Deep Neural Networks (DNNs) have significantly impacted the field of image compression by achieving remarkable gains in compress efficiency, as evidenced by numerous seminal studies [1], [4]. A foundational contribution was proposed by Balle et al. [2] in 2017, who introduced an image compression algorithm based on deep learning models. They employed VAE to map the input image into a latent space. A parameterized prior is then used to model the probability distribution of these quantized latent variables. Building upon [1] and [2], Minnen et al. [4] employ PixelCNN to establish a novel context model. When integrated with a hyperprior structure, this context model allows for the iterative generation of Gaussian priors for both mean and variance at each spatial location within the latent representation. Based on prior work, a variety of different frameworks and algorithms have emerged. Rhee et al. [5] introduced a novel framework using a coarse-to-fine encoding strategy that aims to isolate and leverage low-frequency components to optimize performance in high-frequency regions. Toderici et al. [6] utilized Recurrent Neural Networks (RNNs) in both encoder and decoder designs, achieving variable-rate image compression. In addition to VAE-based approaches, generative models such as Generative Adversarial Networks (GANs) [7] and diffusion models [8] have been explored. While these techniques excel in delivering high perceptual quality at extremely low bit rates, they may compromise the semantic fidelity of the reconstructed images.

B. Multi-View Image Compression

The fundamental principle behind multi-view compression techniques lies in exploiting the correlation between images to eliminate redundant information more effectively. Traditional video encoding frameworks such as H.264 and HEVC are fundamentally designed for single-view video compression. These architectures can be adapted for multi-view image compression paradigms like Multiview Video Coding (MVC) and Multiview High Efficiency Video Coding (MV-HEVC). The field of deep learning-based stereo image compression was initially explored by Liu et al. [9], who introduced the Deep Stereo Image Compression (DSIC) model. DSIC uses multiple skip connections between discrete encoders and decoders for the left and right views, aiming to enable viewpoint transformation and inter-view information exchange. Additionally, in the DSIC model, a conditional entropy model is employed to capture the dependencies that exist between the image pairs. However, the model suffers from increased computational complexity due to the extensive use of skip connections. Subsequent works like SASIC [10] and MASIC [11] have also focused on viewpoint transformation. SASIC employs the H-matrix, while MASIC utilizes inter-view Mean Square Error (MSE) as the intermediate variable to achieve viewpoint transformation. These models demonstrate notable bit-rate reductions but do require the image pairs to be available during joint encoding, unlike distributed image compression methods. Ayzik et al. [3] introduced DSIN, the pioneering learning-based distributed image compression model. This model segments the side information image into patches to make it match the viewpoint of the primary image and subsequently incorporates it into the decoding process of the primary image. Recently, Mital et al. [12] proposed a new distributed model that features a cross-attention mechanism. This model enhances compression efficiency by aligning features of image pairs during decoding, thereby providing a novel perspective on leveraging internal feature correlations for image compression.

III. Method

In this section, we provide an overview of the CVCDIC model, covering its framework and modules, and then discuss the strategies used for training and testing.

A. Framework

The architecture of the proposed CVCDIC model is depicted in Fig. 1. In the architecture, side information (i.e., the right image) is accessible solely during the decoding phase. The encoder $Encoder_X$ is solely responsible for processing the primary image (i.e., the left image) and comprises a series of convolutional layers followed by Generalized Divisive Normalization (GDN) layers. For feature alignment between the primary and side images, dual encoders with identical architecture are situated at the decoding end. $Encoder_Y$ is used for extracting features from the side information image, while $Encoder_W$ is used for extracting Wyner common information which is used to model dependence between the image pairs [13]. To address the constraint of traditional quantization methods that do not allow for gradient back-propagation, we introduce uniform random noise within the range of [-0.5, 0.5] during the training phase as a substitute. Our model adopts a univariate, non-parametric, fully factorized density function to model the latent representation's probability distribution, similar to the approach in [2]. The decoder consists of a feature extraction network, a multi-level cross view attention module, and a spatial refinement module. These modules work in conjunction to extract and fuse relevant information from the primary and side images. Further details about these modules will be elaborated in subsequent sections.

B. Feature Extraction Network

As depicted in Fig. 1, we integrate multiple feature extraction networks into the decoders for both main and side information images to extract internal features individually. Initially, within the feature extraction network, a convolutional layer with a 1×1 kernel is employed to reduce the input feature channels. Following this, we introduce a channel attention module designed to capture the internal features of the images, the structure of which is detailed in

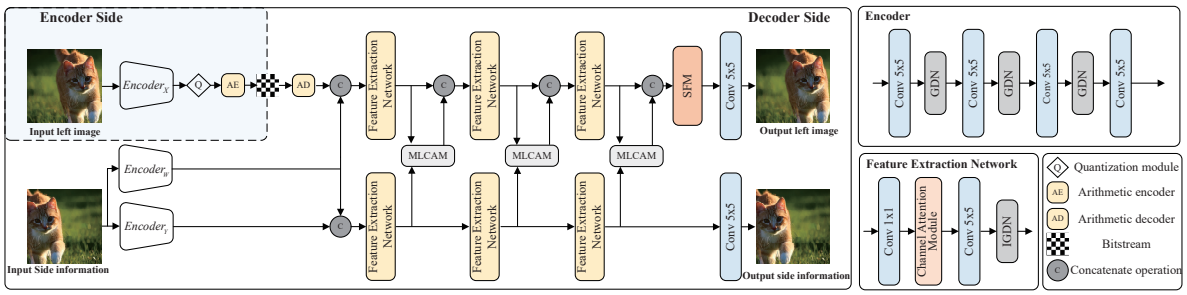

Fig. 1. The framework of the proposed method.

Fig. 2(a). These extracted features are then upscaled by a factor of 2 using a deconvolutional layer with a 5×5 kernel and subsequently activated through an Inverse Generalized Divisive Normalization (IGDN) layer. As detailed in Fig. 2(a), the channel attention module comprises two similar residual structures. Within this module, the input features are sequentially processed through a layer normalization layer, a 1×1 convolutional layer, and a 3×3 depthwise convolutional layer, thereby increasing the number of input feature channels. Inspired by [14], we utilize a Simple Gate mechanism for non-linear activation.

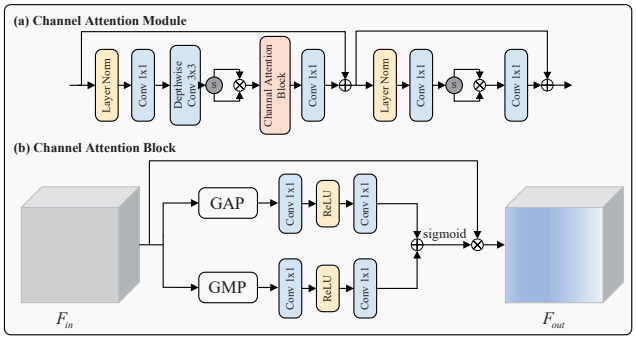

Fig. 2. Illustration of the channel attention module.

To enhance the feature representation of the input, we subsequently design a channel attention block, as illustrated in Fig. 2(b). Two global pooling operations are performed on the input feature F_{in} along the spatial dimension, one with a Global Max Pooling (GMP) layer and another with Global Average Pooling (GAP). The pooled features are separately passed through two sequential 1×1 convolutional layers to obtain two different descriptors. These descriptors are subsequently element-wise summed and activated through a sigmoid function to generate channel-specific weights, which are then used to element-wise multiplied with the input features, resulting in:

$$F_{out} = \sigma\left([\mathrm{conv}(\mathrm{GAP}\left(F_{in}\right)) \oplus \mathrm{conv}(\mathrm{GMP}\left(F_{in}\right))]\right) \otimes F_{in}, \quad (1)$$

where the conv operation involves a sequence of three steps, as illustrated in Fig. 2(b): a 1×1 convolution, followed by ReLU activation, and another 1×1 convolution. Here, $\oplus$ stands for element-wise addition, and $\otimes$ denotes element-wise multiplication. σ denotes the sigmoid nonlinear activation function. Finally, the output of the channel attention block undergoes convolution and is then element-wise added to the module's input to facilitate feature fusion.

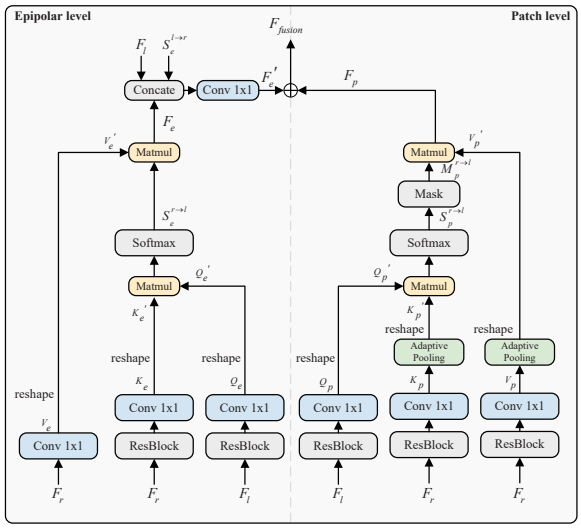

Fig. 3. Illustration of the multi-level cross view attention module.

C. Multi-Level Cross View Attention Module

To address the limitations of existing algorithms like SASIC, which primarily focus on matching image pairs along their epipolar lines, we introduce a multi-level cross view attention module for more deep cross-view information fusion. Unlike traditional methods, our module captures not only epipolar correlations but also those existing at other spatial positions in the side images, enhancing image reconstruction during the decoding process. As illustrated in Fig. 3, our module adaptively pools features from both the main and auxiliary views according to their spatial dimensions, and then explores patch-level correlations between images. This approach synergizes with traditional epipolar line-based feature matching, offering a comprehensive solution for cross-view feature fusion.

In detail, as shown in Fig. 3, during the feature fusion process at the epipolar line level, features from the left image $F_l \in \mathbb{R}^{B\times C\times H\times W}$ and the right image $F_r \in \mathbb{R}^{B\times C\times H\times W}$ pare processed through residual and convolutional blocks to generate a query $Q_e \in \mathbb{R}^{B\times C\times H\times W}$ and a key $K_e \in \mathbb{R}^{B\times C\times H\times W}$, respectively. These are reshaped into $Q'_e \in \mathbb{R}^{(B\times H)\times W\times C}$ and $K'_e \in \mathbb{R}^{(B\times H)\times W\times C}$ to calculate the correlation matrix, which represents the relationship between each column of information in the left and right views along the epipolar line. This matrix is then normalized using a softmax function to generate $S_e^{r\to l} \in \mathbb{R}^{(B\times H)\times W\times W}$. By multiplying $S_e^{r\to l}$ with value matrix $V'_e \in \mathbb{R}^{(B\times H)\times W\times C}$,

we can obtain the weighted right view feature F_e. The above process can be written as

$$F_e = \text{softmax}(Q'_e {K'_e}^{\mathrm{T}}) V'_e. \tag{2}$$

Finally, F_e, F_l, and $S_e^{l \to r}$ (the counterpart of $S_e^{\to l}$) are concatenated along the channel dimension. These concatenated features are then processed through a convolutional layer to yield F'_e, which serves as the output for inter-view feature fusion of the epipolar level.

Our method is not limited to epipolar level information fusion; it also explores inter-image correlations at the patch level. Using a similar approach to the one employed at the epipolar level, we generate Q_p, K_p, and V_p matrices for patch-level analysis. We adopt adaptive pooling operation to K_p, and V_p to obtain pooled features with dimensions $B \times C \times H' \times W'$ and then reshape it to $B \times C \times (H' \times W')$. The patch-level correlation matrix $S_p^{r \to l} \in \mathbb{R}^{B \times (H \times W) \times (H' \times W')}$ can be obtained using

$$S_p^{r \to l} = \text{softmax}({Q'_p}^{\mathrm{T}} K'_p), \tag{3}$$

where $Q'_p \in \mathbb{R}^{B \times C \times (H \times W)}$ and $K'_p \in \mathbb{R}^{B \times C \times (H' \times W')}$.

The correlation matrix $S_p^{r \to l}$ does not directly impact $V'_p \in \mathbb{R}^{B \times (H' \times W') \times C}$ at the patch level. Specifically, a masking operation is applied to $S_p^{r \to l}$. Values in $S_p^{r \to l}$ that are less than τ are set to zero, aiming to minimize the influence of irrelevant information. The masking operation can be formally described as

$$M_p^{r \to l}(i,j) = \begin{cases} S_p^{r \to l}(i,j), & \text{if } S_p^{r \to l}(i,j) > \tau \\ 0, & \text{otherwise} \end{cases}. \tag{4}$$

The final result of inter-view feature fusion along the patch level is computed as

$$F_p = M_p^{r \to l} V'_p. \tag{5}$$

To sum up, the output of the multi-level cross view attention module can be fully characterized by combining the inter-view feature fusions along both the epipolar and patch levels:

$$F_{fusion} = F'_e V'_p \oplus F_p. \tag{6}$$

D. Spatial Refinement Module

To further enhance the quality of image reconstruction, we introduce a spatial refinement module, which consists of an information multi-distillation structure and a spatial attention block, as depicted in Fig. 4. Initially, the fused features outputted by the last cross view attention module are processed through a 1×1 convolutional layer in the SRM for channel reduction. Following this, the features are fed into the information multi-distillation structure. Each distillation structure comprises a convolutional layer followed by a LeakyReLU layer, ending with a channel split operation. As shown in Fig. 4, the output from each distillation structure is divided into two feature sets, F_{si} and F_{ri} (where $i = 1, 2, 3$), where F_{si} is forwarded to the subsequent distillation structure in the sequence. By concatenating F_{r1}, F_{r2}, F_{r3}, and F'_{s3} along the channel dimension, F_{sab} can be obtained, which subsequently processed by the spatial attention module.

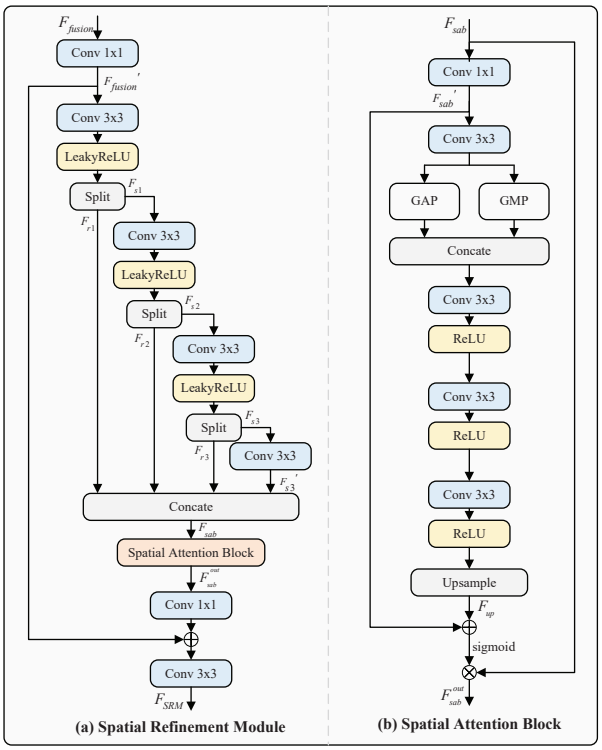

Fig. 4. Illustration of the spatial refinement module.

The spatial attention block aims to dynamically allocate the significance of each spatial location within the feature representation to enhance feature extraction. In the forward pass, the input feature F_{sab} first undergoes channel reduction via a 1×1 convolutional layer generate F'_{sab}. F'_{sab} is then spatially downsampled using a 3×3 convolutional layer. Following this, max and average pooling operations along the channel dimension are executed on the downsampled feature. After concatenating the two features obtained through pooling, this feature is processed through three convolutional layers with ReLU activation functions. The processed features are then upsampled and concatenated with F'_{sab}. The output is sigmoid-activated and multiplied element-wise with the F_{sab}. The formal mathematical description of this process can be written as

$$F_{sab}^{out} = \text{sigmoid}(F'_{sab} \oplus F_{up}) \otimes F_{sab}. \tag{7}$$

The final output of the spatial refinement module can be expressed as

$$F_{SRM} = conv_3(conv_1(F_{sab}^{out}) \oplus F'_{fussion}), \tag{8}$$

where $conv_3$ and $conv_1$ denote convolutional layers with 3×3 and 1×1 kernel, respectively.

E. Training Strategy

We trained our model using the RD loss function. Within this loss function, we incorporated two hyperparameters, α and β, to modulate the relative importance of side information loss and Wyner common information loss, respectively. The mathematical definition of the total loss function is defined as

$$L = (R_x + \lambda D_x) + \alpha (R_y + \lambda D_y) + \beta R_w, \tag{9}$$

where R_x and D_x are the bitrate and distortion for the primary image, R_y and D_y are the bitrate and distortion for the side information, R_w represents the Wyner common information loss, and λ is another hyperparameter that balances the trade-off between bitrate R and distortion D.

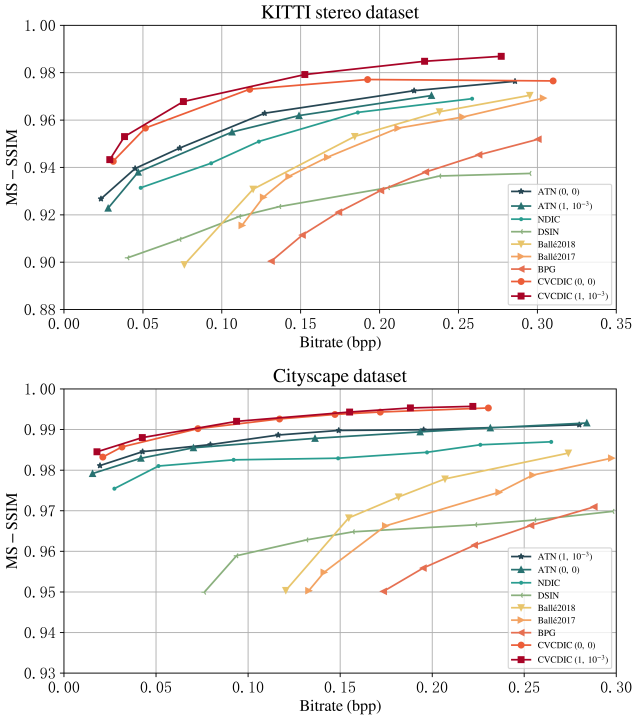

Fig. 5. Plot of CVCDIC against various compression baselines.

IV. Experiment

In this section, we conduct a series of experiments to evaluate the performance of CVCDIC against several learning-based algorithms, such as NDIC [13] and ATN [12]. To ensure a fair comparison in our experiments, we followed the same training settings as those employed in [12]. We implemented our model using the PyTorch framework [15] and conducted experiments on both the KITTI Stereo [16] and Cityscape [17] datasets. For the KITTI Stereo dataset, we used 1576 image pairs for training and allocated 790 pairs each for the validation and test sets. Initially, we resized the images to 370×740 pixels through center cropping and then downsampled them to 128×256 pixels for training. Regarding the Cityscape dataset, we included 2975, 500, and 1525 image pairs in the training, validation, and test sets, respectively. These images were directly downscaled to 128×256 pixels. Optimization was performed using the AMSGrad optimizer [18], with an initial learning rate of 0.0001. To prevent the model from being trapped in local optima, we implemented an adaptive learning rate strategy that reduced the learning rate by a factor of 10 whenever the loss function plateaued, with the minimum learning rate set to 1×10^{-7}. The parameter τ in Eq. (4) is set as 0.009.

A. Results and Analysis

1) Objective evaluation: To evaluate the effectiveness of the CVCDIC model, we conducted comprehensive experiments and compared it with both traditional algorithms like PNG and learning-based approaches, including NDIC [13], DSIN [3], and the state-of-the-art ATN [12]. Following the methodology outlined by Mital et al. [12], we employed the 4:4:4 chroma format for the BPG method and measured compression efficiency using bits per pixel (bpp) and Multi-Scale Structural Similarity (MS-SSIM) as our evaluation metrics. We performed tests across multiple datasets, as expressed by the rate-distortion curves in Fig. 5. The values in parentheses in the legend represent the α and β in Eq. (9), which control the training loss.

In our experiments, carried out on both the KITTI Stereo and Cityscape datasets, CVCDIC consistently outperformed the competing algorithms at both low and high bitrates. On the KITTI Stereo dataset, CVCDIC achieved an MS-SSIM of 0.978 at a bpp of 0.15, surpassing DSIN (0.925), NDIC (0.956), and ATN (0.955) by significant margins. Similarly, on the Cityscape dataset, our model excelled across all bitrates, achieving a 0.005 and 0.011 MS-SSIM improvement over ATN and NDIC, respectively, at 0.15 bpp. The influence of hyperparameters α and β on compression performance was also investigated. Conducted on two different datasets, as shown in Fig. 5, their impact is particularly noticeable in the KITTI Stereo dataset. Specifically, setting $\alpha = 1$ and $\beta = 0.001$ significantly improved performance on both datasets compared to when $\alpha = 0$ and $\beta = 0$. This indicates that incorporating side information penalties related to compression rate and image quality into the loss function enables more effective model optimization.

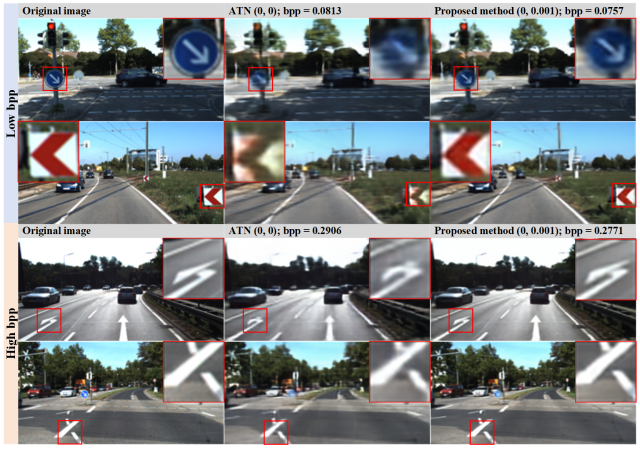

Fig. 6. Visual evaluation of the reconstructed images from the CVCDIC and the ATN.

2) Visual evaluation: We conducted an in-depth evaluation to scrutinize the image compression efficacy of CVCDIC across various bitrates, specifically using the KITTI Stereo datasets. The results are visualized in Fig.6. To provide a nuanced view, we zoomed into specific regions within both the original and reconstructed images. We ensure a fair comparison by compressing the same set of images using both the ATN and CVCDIC algorithms, with closely matched bitrates for each. For this experiment, we selected the results generated by the ATN algorithm configured with $\alpha = 0$ and $\beta = 0$ as it demonstrated superior performance

in KITTI Stereo datasets compared to settings with $\alpha = 1$ and $\beta = 0.001$. As evident in Fig.6, CVCDIC significantly outperforms ATN in terms of image clarity at both low and high bpp settings. Specifically, images reconstructed using CVCDIC display markedly clearer and more detailed road signs compared to those produced by ATN. Furthermore, CVCDIC retains the original colors of road signs, whereas ATN loses color information. As we transitioned to higher bitrate settings, the performance advantages of CVCDIC became even more noticeable. The visibility of road lines in CVCDIC-reconstructed images was substantially superior to that in images generated by ATN. This improvement holds critical practical implications, especially in domains like road analysis and autonomous driving. These results validate the performance of CVCDIC, demonstrating its capability to maintain high-quality image reconstructions across a diverse range of bitrate conditions.

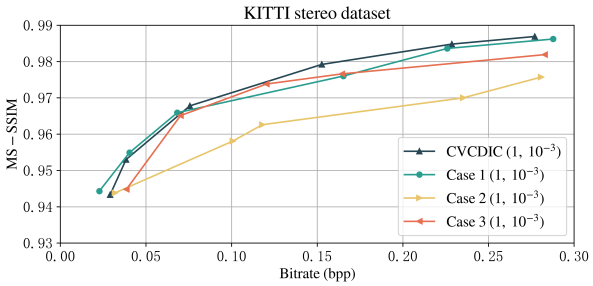

Fig. 7. Rate-distortion curves for ablation study.

3) Ablation study: In this section, we evaluate the contributions of individual modules in CVCDIC through a series of experiments. We separately remove the channel attention module (referred to as Case 1), the multi-level cross view attention module (Case 2), and the spatial refinement module (Case 3) from the complete CVCDIC model. The resulting rate-distortion curves for these configurations are depicted in Fig. 7. Compared to the performance of the complete CVCDIC model, each modified configuration exhibited a decline in effectiveness. This decline is most significant in Case 2, emphasizing the importance of the multi-level cross view attention module in the efficient exchange of inter-image information. The performance degradation in Cases 1 and 3 further emphasizes the importance of effective feature extraction and image reconstruction refinement, respectively. In summary, our ablation study affirms that each module within CVCDIC serves a critical role in optimizing the overall performance of the image compression algorithm.

V. Conclusion

In order to efficiently utilize the side information in the decoder, the CVCDIC is proposed, which excels in both compression efficiency and image reconstruction quality. By integrating a multi-level cross view attention module and a spatial refinement module, CVCDIC leverages and optimizes the fuse efficiency of information during the compression process. The model employs multiple feature extraction networks during decoding to separately process primary and side information images, significantly enhancing compression performance. Additionally, we present a novel multi-level cross view attention module that operates at both the epipolar and patch levels, enabling a comprehensive fuse of information between image pairs. Furthermore, the spatial refinement module adopts information distillation techniques to substantially improve image reconstruction quality. Our experiments, conducted on diverse datasets, confirm that CVCDIC outperforms existing methods such as ATN, NDIC, and DSIN in terms of rate-distortion metrics.

References

[1] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations*, 2018.

[2] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *5th International Conference on Learning Representations*, 2017.

[3] S. Ayzik and S. Avidan, "Deep image compression using decoder side information," in *European Conference on Computer Vision*. Springer, 2020, pp. 699–714.

[4] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *Advances in neural information processing systems*, vol. 31, p. 10771—10780, 2018.

[5] H. Rhee, Y. I. Jang, S. Kim, and N. I. Cho, "LC-FDNet: Learned lossless image compression with frequency decomposition network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6033–6042.

[6] G. Toderici, S. M. O'Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthankar, "Variable rate image compression with recurrent neural networks," *arXiv preprint arXiv:1511.06085*, 2015.

[7] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative adversarial networks for extreme learned image compression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 221–231.

[8] R. Yang and S. Mandt, "Lossy image compression with conditional diffusion models," *arXiv preprint arXiv:2209.06950*, 2022.

[9] J. Liu, S. Wang, and R. Urtasun, "Dsic: Deep stereo image compression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3136–3145.

[10] M. Wödlinger, J. Kotera, J. Xu, and R. Sablatnig, "Sasic: Stereo image compression with latent shifts and stereo attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 661–670.

[11] X. Deng, Y. Deng, R. Yang, W. Yang, R. Timofte, and M. Xu, "MASIC: Deep mask stereo image compression," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[12] N. Mital, E. Özyilkan, A. Garjani, and D. Gündüz, "Neural distributed image compression with cross-attention feature alignment," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2498–2507.

[13] N. Mital, E. Özyılkan, A. Garjani, and D. Gündüz, "Neural distributed image compression using common information," in *Data Compression Conference*, 2022, pp. 182–191.

[14] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *European Conference on Computer Vision*, 2022, pp. 17–33.

[15] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[16] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[17] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[18] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," in *International Conference on Learning Representations*, 2018.