# END-TO-END DEPTH MAP COMPRESSION FRAMEWORK VIA RGB-TO-DEPTH STRUCTURE PRIORS LEARNING

*Minghui Chen*[#] *Pingping Zhang*[†] *Zhuo Chen*[‡] *Yun Zhang*[+] *Xu Wang*[#] *Sam Kwong*[†]

[#] College of Computer Science and Software Engineering, Shenzhen University, China
[†] Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China
[‡]Institute for Infocomm Research, A*STAR, Singapore
[+]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China.

## ABSTRACT

In this paper, we propose a novel framework to exploit and utilize the shared information in RGB-D data for efficient depth map compression. Two main codecs, designed based on existing end-to-end image compression network, are adopted for RGB image compression and enhanced depth image compression with RGB-to-Depth structure prior, respectively. In particular, we propose a Structure Prior Fusion (SPF) module to extract the structure information from both RGB and depth codecs at multi-scale feature levels and fuse the cross-modal feature to generate more efficient structure priors for depth compression. Extensive experiments show that the proposed framework can achieve competitive rate-distortion performance as well as RGB-D task-specific performance at depth map compression compared with the direct compression scheme.

***Index Terms*—** Depth map compression, cross-modal, feature fusion

## 1. INTRODUCTION

Various fields such as 3D reconstruction, autonomous driving, virtual reality, and many visual analysis tasks have developed rapidly with the support of RGB-D data. These practical applications need a massive volume of RGB-D data. With the limitation of storage and bandwidth, the high efficient compression for RGB-D data can reduce the volume of data and improve data processing efficiency.

There are two different modals of data inner the RGB-D image pair. RGB images usually contain rich texture contents and color information, while depth maps are characterized by large smooth regions and sharp edges. In addition, the pixel value of the RGB image represents the intensity of color, while the depth one represents the distance between the camera and the surface of the target object. As shown in Fig. 1 (a), due to the distinct properties of RGB images and depth maps, the direct approach for RGB-D data compression is to compress them independently. There is a range of efficient techniques for RGB image compression including JPEG [1], JPEG2000 [2] and BPG [3]. These traditional codecs are composed of several hand-crafted modules such as transform, quantization and entropy coding, which are designed and optimized independently. Recently, the learning-based image compression method has developed rapidly. Ballé *et al.* [4] propose an end-to-end image compression framework with generalized divisive normalization (GDN) [5], which shows the great capability of redundancy reduction in image compression. After that, Ballé *et al.* [6] further improve the previous work by adopting hyperprior to capture the redundancy in latent as well as modeling the entropy as a conditional Gaussian distribution. Recent studies [7–10] focus on proposing more efficient entropy models to estimate the probability distribution of latent, while others propose more powerful transform architecture such as attention [10,11] and invertible structures [12] to reduce the redundancies.

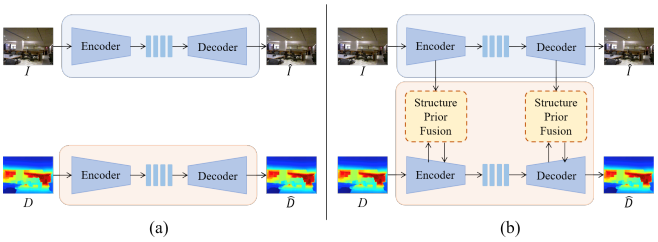


**Fig. 1**. The overview of the traditional RGB-D compression pattern and our proposed framework. (a) Traditional RGB-D compression pattern; (b) Enhanced depth map compression with RGB-to-Depth structure priors.

Various methods adopt the mature RGB codecs for depth map compression. For instance, Pece *et al.* [13] transform the single-channel depth map with high bit-depth into a standard three channels color image to fit the codecs that do not support high bit-depth image. Other methods focus on enhancing compression performance by preprocessing the depth map before compression by RGB codecs. For instance, Fu *et al.* [14] and Panjaitan *et al.* [15] preprocess the depth data to suppress spatial noises and rebuild the depth continuity for efficient coding. However, these methods do not modify the codecs, which are optimized based on RGB images without consid-

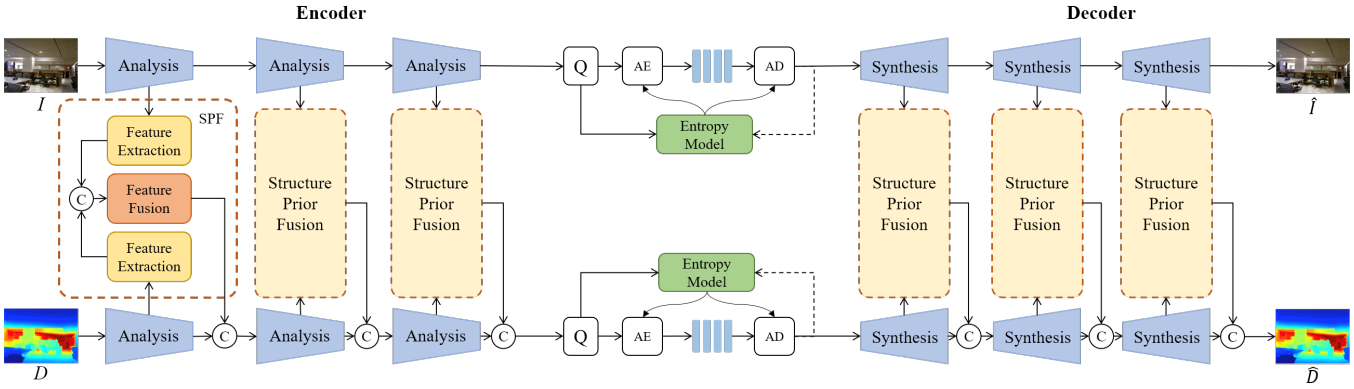


**Fig. 2**. An example architecture of our proposed depth map compression framework with three pairs of SPF modules.

ering the properties of depth maps. Furthermore, some existing learning-based RGB image compression methods can compress the depth map as a gray-scale image with distortion loss designed for depth maps. Although these methods can achieve good performance, compressing RGB images and depth maps separately ignores the redundancy between different modal data, limiting the compression performance.

Another type of approach improves the depth map compression performance with the assistance of RGB information. Farrugia *et al.* [16] extract contour correlation exist between color and depth images for efficient compression. Kazunori *et al.* [17] compressed the depth map with a transformation matrix constructed from the given RGB image. Georgiev *et al.* [18] compress depth map based on depth down-sampling guided by color image segmentation. However, the methods with RGB guidance heavily rely on handcrafted features, which may lead to sub-optimal results.

As shown in Fig. 1 (b), in this paper, we propose a novel framework based on the existing end-to-end image compression framework for RGB-D data to improve depth map compression. Specifically, we exploit and utilize the structure information shared inner RGB-D data to reduce the cross-modal redundancies in the depth map. The main contributions are provided as follows.

- We propose a novel framework for RGB-D data to improve the depth map compression by exploiting the shared information inner RGB-D data.
- We propose a Structure Prior Fusion (SPF) module to efficiently extract and fuse the structure information from two different modals at multi-scale feature levels, which efficiently reduces the redundancies across modals for depth map compression.
- Experimental results show the effectiveness of our proposed framework in terms of rate-distortion measurements and RGB-D task-specific criterion.

## 2. APPROACH

The RGB modal and depth modal exist gaps since their properties are distinct. Meanwhile, they contain some similarities, *e.g.*, some edges of RGB images and depth maps reflect the consistency of semantic information. Therefore, we propose a novel depth map compression framework, which extracts structure priors from both RGB and depth modals and fuses them efficiently to improve depth map compression.

### 2.1. Depth Map Compression Framework

As illustrated in Fig. 2, our proposed depth map compression framework is designed based on the end-to-end image compression architecture, including encoder, decoder and entropy model. The encoder and decoder share symmetric structure, as the encoder consists of analysis modules for RGB images ($I$) and depth maps ($D$), while the decoder contains synthesis modules. Analysis modules serve as the function of the compact latent extraction, while the synthesis modules aim to decode the image signals. To illustrate the flexibility of our framework, analysis modules can be from existing codecs, *e.g.*, Minnen(2018) [7] and Cheng(2020) [10]. Unlike the direct compression pattern that ignores the shared information inner RGB-D data, we further propose SPF modules working on the encoder and decoder, which can exploit structure information from both RGB and depth codecs at multi-scale feature levels to reduce the cross-modal redundancy inner RGB-D pair. Entropy model [6, 7, 10] estimate the probability distributions of the latent for entropy coding.

Training of the compression model is progress in optimizing the following Rate-Distortion cost function $\mathcal{J} = R + \lambda\mathcal{L}_D$, where $R$ is the bitrate approximated by entropy model, $\mathcal{L}_D$ is the distortion measurement between the original and output images. $\lambda$ is the hyper parameter to adjust the rate-distortion trade-off. We adopt the commonly used MSE as distortion loss term for RGB image compression. However, MSE is not suitable for optimizing the depth map compression since MSE causes over-smooth and blurry artifact [19],

leading fidelity degradation in edge regions. Instead, we adopt the following distortion term $\mathcal{L}_D$ introduced in [20]:

$$\mathcal{L}_D = \alpha \times \mathcal{L}_{re} + \lambda_G \times \mathcal{L}_G + \lambda_S \times \mathcal{L}_{SSIM}, \quad (1)$$

where the first term $\mathcal{L}_{re}$ is the pixel-wise $L1$ loss between $D$ and output $\hat{D}$. $\alpha$, $\lambda_G$ and $\lambda_S$ are the weights of different distortion terms, and $\mathcal{L}_G$ is the $L1$ gradients loss defined as:

$$\mathcal{L}_G(D, \hat{D}) = \frac{1}{n} \sum_{\mathbf{p}}^{n} |G_h(D_\mathbf{p}, \hat{D}_\mathbf{p})| + |G_v(D_\mathbf{p}, \hat{D}_\mathbf{p})|, \quad (2)$$

where $n$ is the number of pixels and $G_h$ and $G_v$ compute the gradients differences in the horizontal and vertical directions, respectively. $\mathcal{L}_{SSIM}$ is a modified form of structural similarity index measure (SSIM) [21] defined as $\mathcal{L}_{SSIM}(D, \hat{D}) = \frac{1-\mathbf{SSIM}(D,\hat{D})}{2}$.

### 2.2. RGB-to-Depth Structure Priors Learning

As for the RGB-D pair, due to the high consistency of viewpoints of RGB and depth, the correlation that exists between RGB and depth modal may provide a strong prior to help compression. This motivates us to consider extracting the cross-modal correlation and reducing redundancies in depth map compression.

In order to exploit the structure correlation inner RGB-D pair, we propose a module called SPF for structure prior extraction and fusion. The SPF is embedded at the end of each transform block, namely analysis and synthesis modules. We firstly use the *Feature Extraction* block, which is composed of a $3 \times 3$ convolution layer and an activation layer, to extract structure information from the RGB latent features. The same operation is utilized in the depth codec to extract guidance features for the following feature fusion. Then the structure feature maps of two modals will be concatenated. In the *Feature Fusion* block, the preliminary context is prepossessed by a channel-reduction $3 \times 3$ convolution and an activation layer then reweighted with an attention-based ESA block [22]. Then, the structure features from different modals can be effectively selected and fused according to the importance of content learned by the model via *Feature Fusion* block. Finally, the reweighted features are transferred back to depth codec as redundancy for reduction.

## 3. EXPERIMENTS

### 3.1. Dataset

All the models are trained and tested over the widely used NYU Depth Dataset V2 (NYUv2) [23]. The color-depth pairs, with a size of $640 \times 480$, are captured from 464 scenes via a Microsoft Kinect. We randomly select a subset which contains 4800 pairs for training and 200 pairs for validation. We use the standard test set, including 654 pairs. All the depth maps are inpainted to fill the missing values as mentioned in [23] and evaluated with valid values by adopting a mask during testing following the setting in [24]. We also augment the training data with a random crop of size $256 \times 256$ as well as random horizontal and vertical flips.

### 3.2. Implementation

In the following experiments, we implement the proposed framework based on *mbt2018* [7] and *cheng2020-attn* [10] models provided by CompressAI [25], denoted as *Minnen+SPF* and *Cheng+SPF*. Details of architecture are presented at supplemental materials[1]. There are two stages for model training. We initially train the RGB codec for both *mbt2018* and *cheng2020-attn* at a certain bitrate. Then we freeze the parameters of the RGB codec and further train the SPF modules and the depth codec. For both stages, the models are optimized using Adam optimizer with the initial learning rate of $10^{-4}$ in the first 150 epochs and reducing to $10^{-5}$ for 50 epochs. In addition, $\alpha$, $\lambda_G$ and $\lambda_S$ is empirically set at 0.1, 1 and 1, respectively. The hyper parameter $\lambda$ is formulated as $\lambda = 255^2 \times \lambda_1$. Specifically, we train a high and low rate RGB model by setting $\lambda_1$ as 0.05 and 0.005, respectively. For each RGB model, $\lambda_1$ is set from 0.0001 to 0.001 to meet various bitrate settings during training the depth codec.

### 3.3. Evaluation

To measure the distortion of depth map compression, we use the peak signal-to-noise ratio (PSNR) as well as the standard six metrics used in depth estimation method [26], including average relative error (REL), root mean squared error (RMS), average $\log_{10}$ error and threshold accuracy ($\delta$). Specially, we set the threshold to 1.02 to demonstrate the distinction among different methods more clearly. For standard six metrics evaluation, we scale the depth value to the real depth range from $0\ m$ to $10\ m$. In addition, we also evaluate the compression performance by the RGB-D task-specific metric. We adopt ESANet [27] for RGB-D semantic segmentation and calculate the segmentation metric mean intersection over union (mIoU) for quantitative evaluation. We follow the 40-class label setting of NYUv2 dataset and depth value normalization as mentioned in [27]. Bits per pixel (bpp) is adopted as the bitrate measurement. For PSNR, mIoU and threshold accuracy metrics, the higher value indicates the better performance. For the other distortion metrics, the lower value represents the better result.

We compare our proposed methods with both traditional codecs, including VVC [28], BPG [3] and JPEG2000 [2], and learning-based methods, including Cheng *et al.* [10] and Minnen *et al.* [7]. The caparison methods follow the pattern of direct compression, which is lack of SPF modules and ignores the cross-modal redundancies inner RGB-D data. For the traditional codecs, JPEG2000 is adopted for 16-bit high dynamic

[1] https://github.com/mingfaichen/r2dcompression

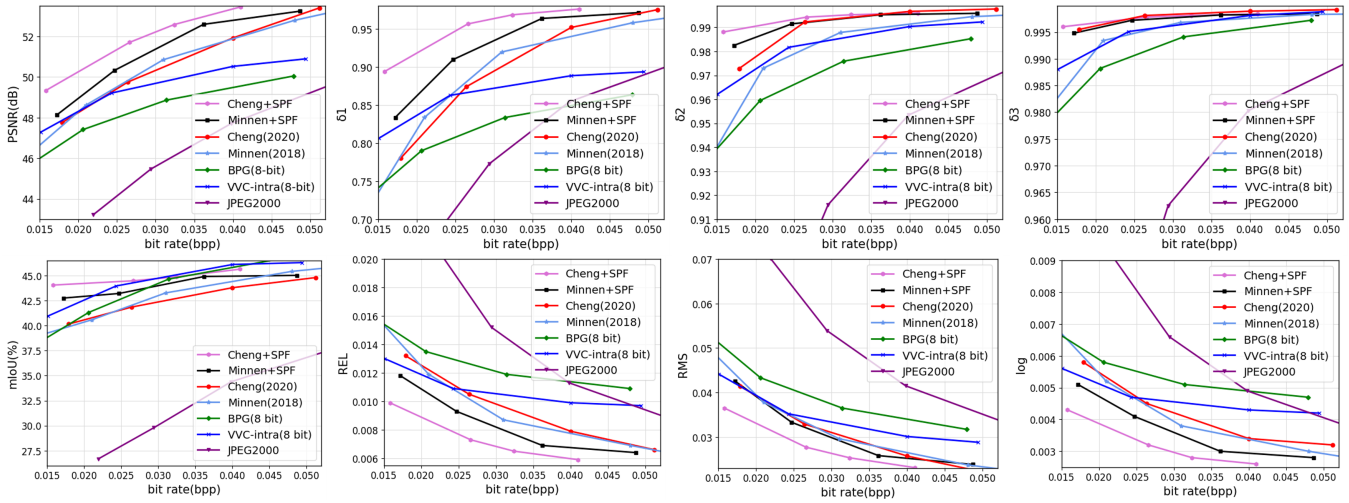


**Fig. 3**. Quantitative evaluation results. The experiments are conducted under lossy RGB images with average bpp 0.6 and average PSNR 38 dB.

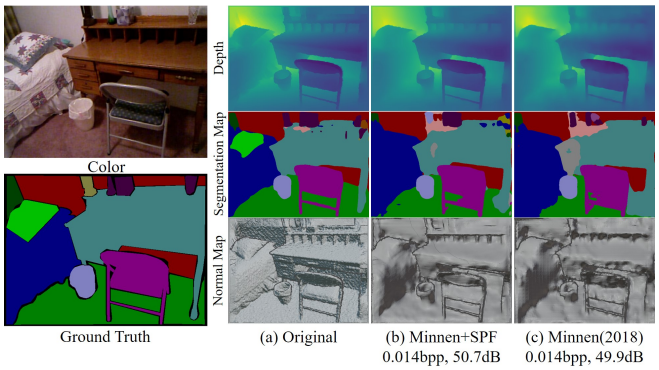


**Fig. 4**. Comparison of visualization results. The first column shows the lossy RGB image compressed at certain bitrate and the ground truth segmentation result. The second to fourth columns show the RGB-D segmentation results and normal maps generated from original and compressed depth maps.

range images, while BPG and VVC are for 8-bit coding due to their limitation of the bit-depth support. Specially, we use the intra configuration provided by the VVC Test model (VTM) to compress the single depth map.

### 3.4. Experimental Results

Fig. 3 shows the results of experiments conducted with high rate RGB model. It's observed that our proposed framework outperforms other comparison methods at most of the evaluation metrics such as PSNR, $\delta_1$, REL, RMS and $\log_{10}$ error. In addition, Table 1 illustrates that our proposed framework achieves 0.78 dB and 1.87 dB BD-PSNR gain as well as 14.01% and 31.02% BD-Rate reduction in *Minnen* and *Cheng*, respectively. This demonstrates that our proposed framework can efficiently reduce the cross-modal redundancy in the depth map compression while maintaining fidelity during reconstruction. Regarding the task-specific metric mIoU, our method surpasses other learning-based methods without SPF modules at low bitrate. Since depth maps provide geometric information to RGB images in RGB-D segmentation, revealing the great importance of structure information of depth maps in semantic segmentation. Our proposed SPF module efficiently extracts complementary structure information from the RGB codec, thus successfully preventing structure information from being severely destroyed during compression at a low bitrate.

We further provide some visualization results for qualitative performance comparison. Fig. 4 shows the semantic segmentation result as well as the normal map generated from the compressed RGB-D pair. The depth map compressed by our proposed method obtains more precise segmentation at object such as the bed, floor, chair and table. On the other hand, the normal maps illustrate that our proposed method can preserve more details at the boundaries of the objects. More results are provided in supplemental materials.

**Table 1**. The BD-PSNR and BD-BR metrics between our proposed framework and comparison method.

| Comparison Method | BD-PSNR(dB) | BD-Rate(%) |
|---|---|---|
| Minnen + SPF vs. Minnen(2018) | 0.78 | -14.01 |
| Cheng + SPF vs. Cheng(2020) | 1.87 | -31.02 |

## 4. CONCLUSION

In this paper, we propose a novel framework for depth map compression. The SPF module is designed to extract and fuse cross-modal structure information at multi-scale feature levels. Experimental results indicate that our proposed method effectively reduces the cross-modal redundancy and achieves promising compression performance compared with direct compression pattern.

## 5. REFERENCES


[1] Gregory K Wallace, "The jpeg still picture compression standard," *IEEE transactions on consumer electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

[2] Majid Rabbani and Rajan Joshi, "An overview of the jpeg 2000 still image compression standard," *Signal processing: Image communication*, vol. 17, no. 1, pp. 3–48, 2002.

[3] "Better portable graphics," [EB/OL], https://bellard.org/bpg/.

[4] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli, "End-to-end optimized image compression," in *5th International Conference on Learning Representations (ICLR)*, 2017.

[5] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli, "End-to-end optimization of nonlinear transform codes for perceptual quality," in *2016 Picture Coding Symposium (PCS)*. 2016, pp. 1–5, IEEE.

[6] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston, "Variational image compression with a scale hyperprior," in *6th International Conference on Learning Representations (ICLR)*, 2018.

[7] David Minnen, Johannes Ballé, and George Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Advances in Neural Information Processing Systems 31*, 2018, pp. 10794–10803.

[8] Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack, "Context-adaptive entropy model for end-to-end optimized image compression," in *7th International Conference on Learning Representations (ICLR)*, 2019.

[9] Yueyu Hu, Wenhan Yang, and Jiaying Liu, "Coarse-to-fine hyper-prior modeling for learned image compression," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 11013–11020.

[10] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7936–7945.

[11] Lei Zhou, Zhenhong Sun, Xiangji Wu, and Junmin Wu, "End-to-end optimized image compression with attention mechanism," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[12] Yueqi Xie, Ka Leong Cheng, and Qifeng Chen, "Enhanced invertible encoding for learned image compression," in *MM '21: ACM Multimedia Conference*, 2021, pp. 162–170.

[13] Fabrizio Pece, Jan Kautz, and Tim Weyrich, "Adapting standard video codecs for depth streaming," in *JVRC11: Joint Virtual Reality Conference of EGVE*, Sabine Coquillart, Anthony Steed, and Greg Welch, Eds., 2011, pp. 59–66.

[14] Jingjing Fu, Dan Miao, Weiren Yu, Shiqi Wang, Yan Lu, and Shipeng Li, "Kinect-like depth data compression," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1340–1352, 2013.

[15] Christin Panjaitan, Chung-An Shen, and Shanq-Jang Ruan, "A low complexity depth map compression approach for microsoft kinect devices," in *IEEE 4th Global Conference on Consumer Electronics (GCCE)*, 2015, pp. 322–323.

[16] Reuben A. Farrugia and Isabel Gambin, "Exploiting color-depth image correlation to improve depth map compression," in *Proceedings of Eurocon 2013, International Conference on Computer as a Tool*, 2013, pp. 1676–1683.

[17] Kazunori Uruma, Katsumi Konishi, Tomohiro Takahashi, and Toshihiro Furukawa, "Depth image coding algorithm via the colorization based image coding," in *2016 IEEE 59th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2016, pp. 1–4.

[18] Mihail Georgiev, Evgeny Belyaev, and Atanas P. Gotchev, "Depth map compression using color-driven isotropic segmentation and regularised reconstruction," in *2015 Data Compression Conference(DCC)*, 2015, pp. 153–162.

[19] Yixin Gao, Yaojun Wu, Zongyu Guo, Zhizheng Zhang, and Zhibo Chen, "Perceptual friendly variable rate image compression," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2021, pp. 1916–1920.

[20] Ibraheem Alhashim and Peter Wonka, "High quality monocular depth estimation via transfer learning," *CoRR*, vol. abs/1812.11941, 2018.

[21] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[22] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu, "Residual feature aggregation network for image super-resolution," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2356–2365.

[23] Nathan Silberman, Pushmeet Kohli, Derek Hoiem, and Rob Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV 2012*, 2012.

[24] David Eigen, Christian Puhrsch, and Rob Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 2366–2374.

[25] 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.

[26] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, 2016.

[27] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross, "Efficient RGB-D semantic segmentation for indoor scene analysis," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13525–13531.

[28] "Vvc official test model vtm," [EB/OL], https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM.