Learning based Multi-modality Image and Video Compression
(Supplementary Materials)

Due to the space limitation in the main paper, we provide more implementation details and comprehensive results in the supplementary material.

1. Experiments

Multi-modality Image Compression on the FLIR Dataset
In the main paper, we only provide the BDBR results for the compression results on the FLIR dataset. Here, Fig. S1 and Fig. S2 further show the rate-distortion curves from different compression approaches for visible and infrared image compression on the FLIR dataset [2]. Compared with the separately optimized single-modality compression method [9], our approach achieves 0.4dB and 0.2dB gains on the FLIR dataset for the visible image compression and infrared image compression, respectively.

Multi-modality Image Compression LPIPS metrics

<table>
<thead>
<tr>
<th>Methods</th>
<th>FLIR visible</th>
<th>FLIR infrared</th>
<th>KAIST visible</th>
<th>KAIST infrared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>-30.560</td>
<td>-19.523</td>
<td>-17.974</td>
<td>-3.891</td>
</tr>
</tbody>
</table>

We evaluate the compression performance in terms of the LPIPS [10] metric. As shown in Fig. S3, our approach obviously outperforms the single-modality approaches [1, 9] for both visible and infrared image compression on both FLIR and KAIST datasets. When compared with the single-modality approach [9], our approach achieves more than 15.8% bitrate savings for visible image compression on the FLIR dataset. The BDBR results are provided in Table S1.

Figure S1. Visible image compression results from different approaches on the FLIR dataset in terms of PSNR, MS-SSIM and FID.

Figure S2. Infrared image compression results from different approaches on the FLIR dataset in terms of PSNR, MS-SSIM and FID.
Figure S3. Visible and infrared image compression LPIPS results from different approaches on the FLIR and KAIST dataset.

Figure S4. RD curves of the proposed framework with element-wise alignment module Ours(EA) and channel-wise alignment module Ours(CA).

**Element-wise Alignment** As shown Fig. S4, we provide the RD curve when our approach using the element-wise alignment module (Ours(EA)). Experimental results show that it performs much worse than our proposed approach with channel-wise alignment (Ours(CA)).

**The number of Spatial Alignment module** In our default implementation, we use 3 spatial alignment modules (Ours(SA)) at the decoder side in the proposed framework and achieve 0.3db improvements over the baseline method [9]. Here we also provide the comparison results using single spatial alignment module (Ours(SA-1)) and 6 spatial alignment modules at both the encoder and decoder sides (Ours(SA-6)). As the shown Fig. S5, the performance improves while increasing the number of modules. The approaches Ours(SA-6) and Ours(SA-1) have 0.4 and 0.2 gains
2. Implementation Details

**Joint Optimization** In the training stage for the joint optimization of infrared and visible image compression, to maintain the performance of the infrared image compression and improve the compression for visible image compression, we additionally introduce a pretrained infrared image model as the reference model. During the training stage, the parameters of the reference model are frozen. Then, the multi-modality compression framework (both infrared and visible image compression) is optimized by using the following rate-distortion loss function,

\[ \mathcal{L}_{RD} = \mathcal{L}^v_{RD} + w |\mathcal{L}^i_{RD} - \mathcal{L}^{i*}_{RD}| \]  

where \( \mathcal{L}^v_{RD} \) represents the loss of the reference infrared model, and \( w \) is a trade-off parameter and set as 5. \( \mathcal{L}^i_{RD} \) and \( \mathcal{L}^{i*}_{RD} \) represent the losses for the visible and infrared image compression in our multi-modality compression. We have reported the experimental results in the main paper and it is observed that the joint optimization only brings marginal performance improvement. Therefore, we do not use the joint optimization in our default implementation.

**Feature Encoder and Feature Decoder** Fig. S7 shows the network structure of the feature encoder and feature decoder in our framework. \( S \) represents the stride of the convolution and deconvolution layer.

**Feature Encoder and Feature Decoder** Fig. S7 shows the network structure of the feature encoder and feature decoder in the multi-modality image and video compression framework. In our approach, the stride \( S \) of the first convolution(deconvolution) layer for the visible modality and infrared modality are set as 2 and 1 in the image compression framework since the size ratio of the color-thermal pairs is 2:1.

**Multi-modality Image Compression Framework** The complete architecture of the multi-modality image compression framework is shown in Fig. S8. We use Minnen’s approach [9] as our baseline to implement our multi-modality image compression framework.

**Spatial Alignment Module** In the spatial alignment module, we set the patch size, window size and the number of heads in the Multi-head Cross Attention (MCA) module as 2, 8 and 3, respectively.

**Multi-modality Video Compression Framework** Fig. S9 shows the complete structure of the multi-modality video compression framework. We use the FVC [6] as the baseline to implement our multi-modality video compression framework. Specifically, we first use FVC to compress the infrared video sequences. For the visible video sequences, in addition to the existing motion compensation in video codec, we further employ the affine transformation to generate more accurate compensated results. Furthermore, we also use the spatial alignment module in the residual decoder for the visible video compression.
Figure S8. The network architecture of our proposed multi-modality image compression. (a) The network architecture for infrared image compression. (b) The network architecture for visible image compression. AE and AD are arithmetic encoder and decoder, respectively, and $N$ is the number of channels and is set as 192 in the experiments. The context model and entropy model follow the design of Minnen's approach [9]. $\hat{x}^i$ and $\hat{g}^i_j \in 1, 2, 3$ in the (b) represent the reconstructed infrared image and intermediate features in the (a).

Figure S9. The complete network structure of multi-modality video compression in Fig. 4 of the main paper.
3. Dataset Description

**FLIR testing dataset** The filenames of the 20 randomly selected thermal-color image pairs from the FLIR dataset [2] are listed below.

- FLIR_08884.png
- FLIR_09042.png
- FLIR_09063.png
- FLIR_09175.png
- FLIR_09218.png
- FLIR_09311.png
- FLIR_09451.png
- FLIR_09673.png
- FLIR_09682.png
- FLIR_09705.png
- FLIR_09706.png
- FLIR_09728.png
- FLIR_09751.png
- FLIR_09792.png
- FLIR_09886.png
- FLIR_09896.png
- FLIR_10082.png
- FLIR_10107.png
- FLIR_10171.png
- FLIR_10217.png

**KAIST testing dataset** The filenames of the 18 thermal-color image pairs from the KAIST dataset [7] are listed below.

- set06/V000/I00000.png
- set06/V001/I00000.png
- set07/V000/I00000.png
- set07/V001/I00000.png
- set07/V002/I00000.png
- set07/V002/I01596.png
- set08/V000/I00000.png
- set08/V001/I00000.png
- set08/V002/I02499.png
- set09/V000/I00000.png
- set10/V000/I00000.png
- set10/V001/I04193.png
- set11/V000/I00000.png
- set11/V001/I02019.png

References