# Learning based Multi-modality Image and Video Compression

Guo Lu[1], Tianxiong Zhong[1], Jing Geng[1], Qiang Hu[2], and Dong Xu[3]

[1]Beijing Institute of Technology, {sdluguo, inkosizhong}@gmail.com, janegeng@bit.edu.cn
[2]ShanghaiTech University, huqiang@shanghaitech.edu.cn
[3]University of Sydney, dong.xu@sydney.edu.au

## Abstract

*Multi-modality (i.e., multi-sensor) data is widely used in various vision tasks for more accurate or robust perception. However, the increased data modalities bring new challenges for data storage and transmission. The existing data compression approaches usually adopt individual codecs for each modality without considering the correlation between different modalities. This work proposes a multi-modality compression framework for infrared and visible image pairs by exploiting the cross-modality redundancy. Specifically, given the image in the reference modality (*e.g.*, the infrared image), we use the channel-wise alignment module to produce the aligned features based on the affine transform. Then the aligned feature is used as the context information for compressing the image in the current modality (*e.g.*, the visible image), and the corresponding affine coefficients are losslessly compressed at negligible cost. Furthermore, we introduce the Transformer-based spatial alignment module to exploit the correlation between the intermediate features in the decoding procedures for different modalities. Our framework is very flexible and easily extended for multi-modality video compression. Experimental results show our proposed framework outperforms the traditional and learning-based single modality compression methods on the FLIR and KAIST datasets.*

## 1. Introduction

In several practical vision applications (*e.g.*, autonomous driving), cameras from different modalities such as visible or infrared imaging cameras are often jointly used for various computer vision tasks by exploiting the complementary characteristic. For example, the visible (RGB) cameras can often provide continuous, high-resolution color images but may not work well for extreme-low lighting scenarios, which is precisely what infrared cameras can help. At the same time, infrared cameras are easily disturbed by abnormal heat sources, but the drawback can be compensated by using visible cameras. However, these multiple modalities visual analysis approaches [10, 13, 25, 26, 47, 48] will increase the storage and transmission costs as more images from different modalities are transmitted to the decoder side for visual analysis. Therefore, how to design an efficient compression method for multi-modality visual data is a new and challenging research problem.

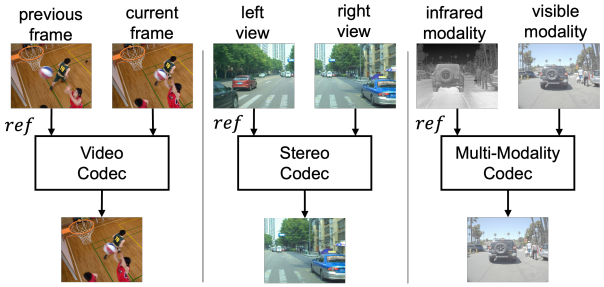


Figure 1. The comparison between video compression, stereo compression and multi-modality compression. Our multi-modality compression approach uses the cross-modality infrared image to facilitate the compression of visible image.

In the past decades, a lot of traditional and learning-based compression methods [1,3,5,6,9,30,32,34,36–38,42] have been proposed for image or video compression. However, most existing works focused on single-modality image compression without considering the correlation between different modalities. Due to the strong correlation between the images from different modalities, we cannot use the existing single-modality compression methods to fully exploit the compression redundancy. One of the most related research topics is stereo image compression, where cross-view redundancy is exploited by using various view alignment approaches. However, compared with the stereo images that share similar distribution, the intensity of different modality images may be quite different (see Fig. 1). Therefore, the commonly used alignment techniques like block-based motion/disparity estimation [19] or homograph transform [14] are not feasible enough for multi-modality compression. Moreover, considering that the multi-modality data like infrared and visible image pairs represent the same scene in different perspectives, the compression for pixel-wise motion/disparity information from most existing estimation approaches will consume a large number of bits for

Jing Geng is the corresponding author.

compression, which is too expensive. Therefore, it is non-trivial to develop a new framework for multi-modality data compression.

In this paper, we propose a learning-based multi-modality compression framework for the infrared and visible image pairs by exploiting the cross-modality redundancy in the feature space. Considering the explicit alignment of different modalities is very difficult and estimated motion/disparity information also requires a lot of transmission bitrates, we use the efficient affine transform and attention mechanism to achieve channel-wise and spatial-wise feature alignment, respectively. Specifically, take the compression procedure of visible image (*i.e.*, RGB image) as an example, based on the extracted features from the decoded infrared and the original visible images, the affine transformation coefficients are estimated, which can be transmitted to the decoder side at marginal bandwidth cost. Then we achieve the channel-wise feature alignment based on the affine transform, and the corresponding transformed features from the infrared modality are used as the conditional context for compressing the visible image. Furthermore, we leverage the correlation of intermediate features from different modalities in the decoding procedure through the spatial-wise alignment module. Our module is integrated into the visible image decoder and will spatially warp the intermediate features from the reference modality to generate the aligned feature, which is used to further reduce the cross-modality redundancy.

The proposed framework is very flexible, and the image from one modality can be easily used as the reference for image compression from another modality. And it can also be easily extended for multi-modality video compression. Experimental results show that the proposed method achieves better compression performance on several benchmark datasets when compared with the single-modality image and video compression approaches. The contributions of our framework are summarized as follows,

- We propose a learning-based framework to compress image pairs from different modalities by exploiting the cross-modality redundancy. As far as we know, it is the first end-to-end optimized framework to compress visible-infrared image pairs.
- Our framework introduces the channel-wise and spatial-wise alignment modules to effectively exploit the correlations between different modalities in the feature space.
- The proposed framework is very flexible and can be extended for multi-modality video compression. Experimental results on several datasets demonstrate the effectiveness of the proposed multi-modality image/video compression framework.

## 2. Related works

**Image and Video Compression.** In the past decades, several representative compression standards [1, 9, 34, 36, 37, 42] are proposed and widely used in many practical applications. Recently, the learning based image and video compression approaches have attracted increasing attention [3–6, 12, 15, 17, 18, 21, 21, 29, 29–33, 33, 38, 39, 44–46] and show comparable or even better performance than the latest image or video compression standards [9, 37]. Although it is feasible to extended these methods for infrared image or video compression [16, 24], the existing standards can only reduce the redundancy in the single modality without exploiting the cross-modality information. Considering the increasing demand for storing and transmitting multi-modality data, like depth map, infrared image or optical flow map, it is necessary to propose a new compression framework for multi-modality data.

**Stereo Image and Video Compression.** Stereo image compression aims to compress a pair of images from different views. To exploit this inter-view redundancy, several multi-view image/video compression standards have been proposed based on the traditional single-view image/video methods, like MV-HEVC [19] or MVC [41]. These approaches use the disparity-based motion compensation [35] to improve the compression performance in addition to the existing inter-frame compensation.

Recent works also try to employ deep neural networks for stereo image compression [14, 27]. Liu *et al.* introduced the parametric skip functions to leverage the disparity-compensated features from the reference view. In [14], the homography matrix is estimated to warp the left view image to the right view image, which reduces the view redundancy. However, these learned stereo image compression approaches are still used for single-modality images recorded by stereoscopic cameras with slightly different positions. The multi-modality data, such as visible and infrared paired images, are captured using different cameras. The internal characteristics of these images are quite different, and existing techniques like homography transform are not feasible. Therefore, it is necessary to develop a multi-modality image compression framework.

**Multi-modality Data Compression.** The multi-modality or multi-sensor information is widely used in various computer vision tasks [10, 13, 25, 26, 47, 48], especially for the 3D vision task. For example, Liang *et al.* [25] utilized both the images and point cloud information to improve 3D object detection accuracy. Zhang *et al.* [47] extract the features from different modalities and fuse these features for object tracking.

In recent years, some multi-modality data compression methods [8, 11, 40] have been proposed. However, these approaches are based on hand-crafted codecs and are mostly designed for multi-view images plus depth images or medi-

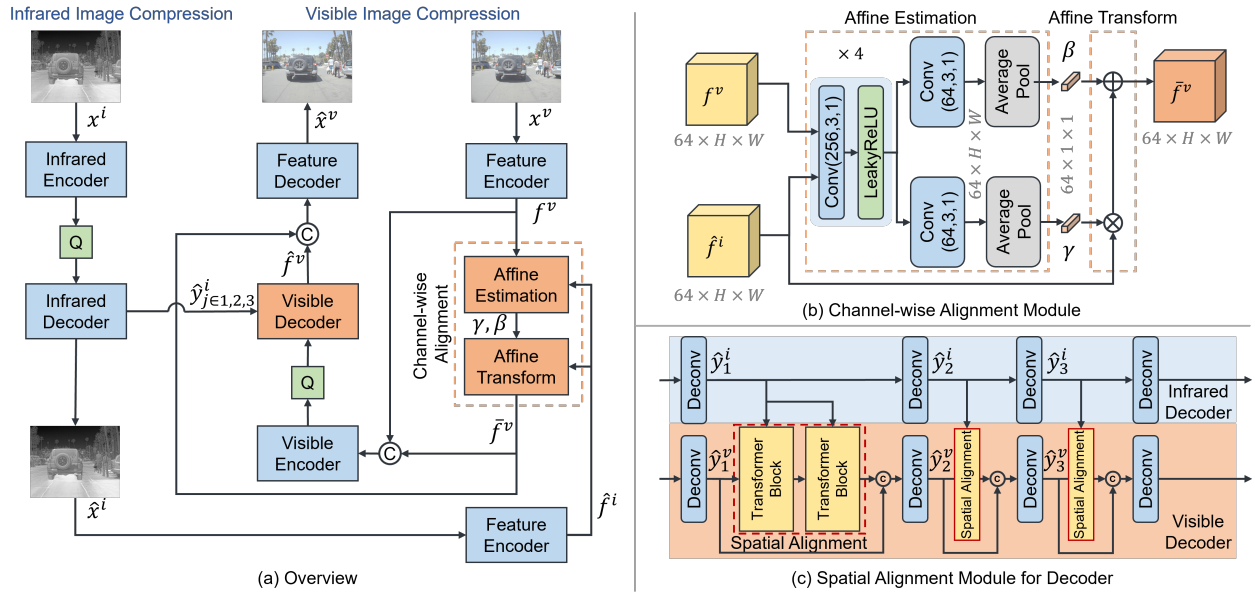


Figure 2. (a) Our multi-modality compression framework, where the decoded infrared image $\hat{x}^i$ is used as the reference to compress the visible image $x^v$. The infrared image is compressed by using the existing image compression method [30]. (b) The network architecture of the channel-wise feature alignment module. 'Conv(C,K,S)' represents the convolution operation with kernel size $K \times K$, stride $S$, and number of output channel as $C$. We use the spatial average pooling layer in our implementation. (c) The illustration of our Transformer-based spatial alignment module in the decoder side. The output feature $\hat{y}_j^i$ from the $j$-th deconvolution layer in the infrared modality is warped to the visible modality and is concatenated with feature $\hat{y}_j^v$ to improve the compression performance.

cal images plus signals. Therefore, the research on the compression of visible-infrared pairs is still blank.

## 3. Method

### 3.1. Overview

The overall architecture of our multi-modality compression approach is shown in Fig.2(a). Here we use the reconstructed infrared image $\hat{x}^i$ as the cross-modality reference to improve the compression performance for the visible image $x^v$. Our approach is flexible and the reconstructed visible image $\hat{x}^v$ can also be used for compressing $x^i$.

As shown in Fig.2(a), we first use the existing image compression method [30] to compress the infrared image $x^i$. Then the features of the visible image $x^v$ and the reconstructed infrared image $\hat{x}^i$ are extracted by the feature encoder module, which is implemented by using several convolution layers. Based on the extracted features, a channel-wise feature alignment module is introduced to calculate the channel-wise affine transformation coefficients $\beta$ and $\gamma$ to align the feature of the infrared modality to the visible modality. In our framework, $\beta$ and $\gamma$ are losslessly transmitted to the decoder side. After that, the aligned feature $\bar{f}^v$ is fed into the visible image encoder network as the context information. Here, we follow network design in [30] to implement the image codec. Finally, the output $\hat{f}^v$ from the visible decoder is concatenated with the aligned feature $\bar{f}^v$ to produce the reconstructed frame $\hat{x}^v$ through the feature decoder.

Considering that the spatial correlation between features in different modalities is not fully utilized in the channel-wise alignment module, we further exploit the correlation between the intermediate features from different modalities by the spatial alignment module in the visible decoder. As shown in Fig.2(c), $\hat{y}_j^i$ and $\hat{y}_j^v$ represent the outputs of the $j$-th deconvolution layers in the infrared and visible decoder, respectively. Our spatial feature alignment module uses the Transformer-based mechanism to spatially warp the intermediate feature from the infrared modality to the visible modality, and the warped feature is used in the decoding procedure. More details are provided Section 3.3. Due to the limited space, we provide the network architectures of feature encoder/decoder and visible/infrared codec (encoder and decoder) in the supplementary material.

The compression network for the visible image is optimized by using the following rate-distortion loss function,

$$\lambda D + R = \lambda d(x^v, \hat{x}^v) + H(\hat{y}^v) + H(\gamma) + H(\beta) \quad (1)$$

where $d(x^v, \hat{x}^v)$ denotes the distortion between the input image $x^v$ and the reconstructed image $\hat{x}^v$. $H(\cdot)$ represents the number of bits used for encoding the representations. In our framework, the latent representation $\hat{y}^v$ is encoded by using the entropy model in [30], and the

channel-wise affine transformation coefficients $\gamma, \beta$ are directly stored and transmitted in Float format at negligible bandwidth cost. $\lambda$ is a hyperparameter used to control the rate-distortion trade-off.

In contrast to the video compression task or the stereo image compression task, the image pairs in multi-modality compression do not share a similar intensity distribution, and the existing alignment approaches like optical flow are not feasible. Therefore, we adopt both channel-wise and spatial-wise feature space alignment approach. Furthermore, considering that the image pairs in different modalities usually represent the same scenario, we only encode the compact affine coefficients $\beta$ and $\gamma$ for a better rate-distortion trade-off.

### 3.2. Channel-wise Alignment Module

In our proposed framework, we use the channel-wise alignment in the feature space to reduce the redundancy between the reconstructed infrared image $\hat{x}^i$ and the visible image $x^v$. The network architecture of our channel-wise alignment module is shown in Fig.2(b). Given the extracted features $f^v$ and $\hat{f}^i$ from the visible image and the reconstructed infrared image, we feed them to several convolutional layers. After that, we use the spatial average pooling to generate the affine transform coefficients $\gamma, \beta \in R^{64\times1\times1}$. Then the feature $\hat{f}^i$ from the decoded infrared image is aligned to the visible modality as follows,

$$\bar{f}^v = \gamma \times \hat{f}^i + \beta \tag{2}$$

where $\times$ and $+$ represents the channel-wise multiplication and addition, respectively. And $\bar{f}^v$ is the aligned feature map. In the encoder side, the aligned feature $\bar{f}^v$ and the original feature $f^v$ are concatenated as the input for the following encoder network. Besides, other alternative solutions like encoding the residual between $\bar{f}^v$ and $f^v$ are also feasible in our framework, and we provide more experimental results in Section 4.4.

At the decoder side, the received affine transform coefficients $\beta$ and $\gamma$ are used to produce the aligned feature $\bar{f}^v$, which will be concatenated with the outputs of the decoder to obtain the final reconstructed frame $\hat{x}^v$ through the feature decoder. Considering that these coefficients are compact, we do not perform any compression, and they are losslessly sent to the decoder side at negligible cost.

### 3.3. Spatial Alignment Module

Since the channel-wise alignment module only exploits the cross-modality redundancy by channel-wise transform, the spatial correlation between features in different modalities is not fully utilized. Our visible decoder uses the spatial feature alignment module to spatially warp the feature from the infrared to the visible modality, based on the similarity between these two features. The whole network architecture

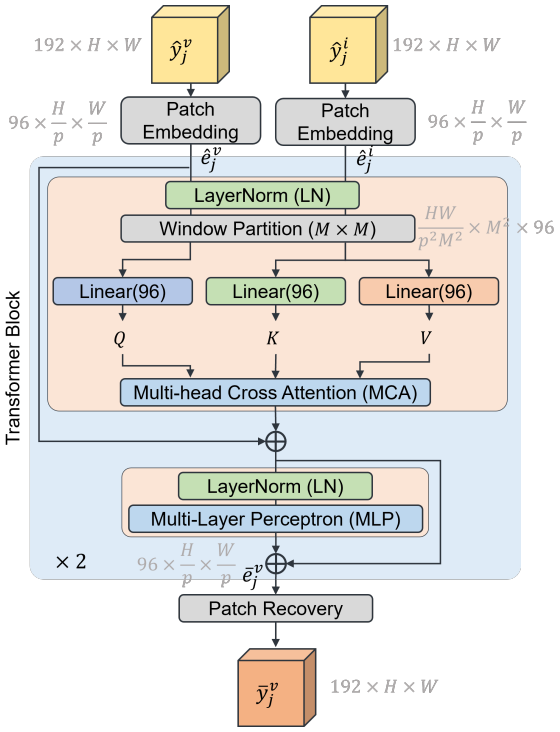


Figure 3. The implementation of our spatial feature alignment module. $\hat{y}_j^v$ and $\hat{y}_j^i$ represent the output intermediate features of the $j$-th deconvolution layers in the decoder for visible image and infrared images. $\bar{y}_j^v$ is the aligned feature from the thermal modality to visible modality. We follow the design in [28] to implement the LayerNorm and MLP networks.

of the spatial alignment module is shown in Fig.3. Inspired by the Swin-Transformer [28], we use the Transformer-based mechanism to exploit the correlation between the intermediate features from the infrared image $x^i$ and the visible image $x^v$ in the decoding procedure.

Specifically, let $\hat{y}_j^i \in R^{192\times H\times W}$ and $\hat{y}_j^v \in R^{192\times H\times W}$ represent the outputs of the $j$-th deconvolution layer of the decoder network for $x^i$ and $x^v$ in Fig.2, respectively. We first perform a $p \times p$ patch embedding operation by using a convolutional layer and generate the corresponding embedding $\hat{e}_j^i \in R^{96\times\frac{H}{p}\times\frac{W}{p}}$ and $\hat{e}_j^v \in R^{96\times\frac{H}{p}\times\frac{W}{p}}$, where $p$ is set as 2. Then, the $\hat{e}_j^i$ and $\hat{e}_j^v$ are fed into the LayerNorm and Multi-head Cross Attention (MCA) module, where the features from different modalities are used to calculate the attention matrices and the infrared embedding $\hat{e}_j^i$ is warped to generate corresponding aligned feature $\tilde{e}_j^v$. After that, we use LayerNorm and MLP networks to further enhance the feature transform [28]. Besides, the residual connection is added to help the training procedure and this Transformer based block is formulated as follows,

$$\begin{aligned} \tilde{e}_j^v &= \mathrm{MCA}(\mathrm{LN}(\hat{e}_j^v), \mathrm{LN}(\hat{e}_j^i)) + \hat{e}_j^v \\ \bar{e}_j^v &= \mathrm{MLP}(\mathrm{LN}(\tilde{e}_j^v)) + \tilde{e}_j^v \end{aligned} \tag{3}$$

In our implementation, we use two Transformer blocks

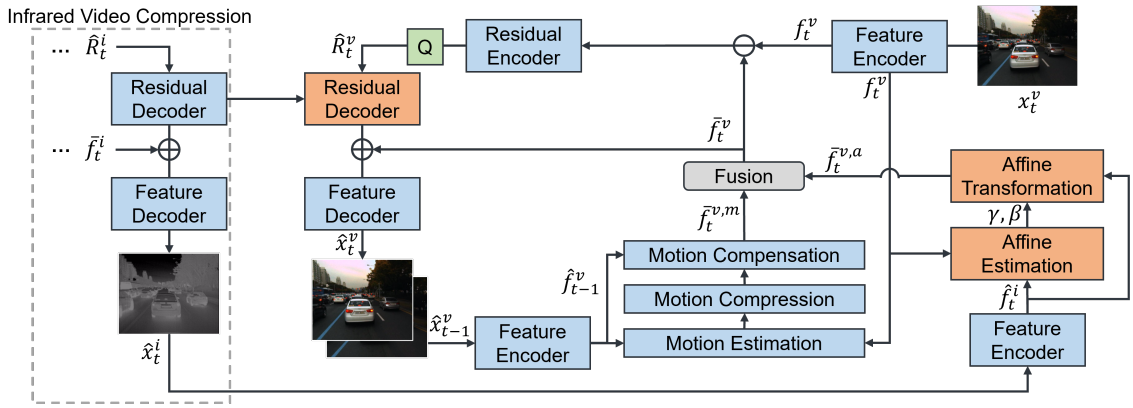


Figure 4. The framework of our Multi-Modality Video Compression Framework. Due to space limitation, we only keep the relevant components like residual decoder for infrared video compression. For visible video compression, the motion compensated feature $\bar{f}_t^{v,m}$ and the channel-wise aligned thermal feature $\bar{f}_t^{v,a}$ are fused together as the predicted feature $\bar{f}_t^v$ to calculate the residual. In addition, we also integrate the spatial alignment module into the residual decoder network for visible video compression and further exploit the cross-modality information from infrared videos.

and the embeddings are shifted before the second Transformer Block. At the end, the generated embedding $\bar{e}_j^v$ is recovered to the intermediate feature $\bar{y}_j^v$ by using a deconvolution layer, which is a reverse process of patch embedding. We use three spatial feature alignment modules on the decoder side, and the output from our module will be fed to the next deconvolutional layer as shown in Fig. 2(c).

**Multi-head Cross Attention.** The multi-head cross attention module will produce the aligned embedding from the infrared image. Specifically, the input embedded features are partitioned into non-overlapping $M \times M$ windows with a shape $\frac{HW}{p^2M^2} \times M^2 \times 96$, where $\frac{HW}{p^2M^2}$ represents the number of windows, and $M$ is set as 4. Then, the module calculates the local attention between the windows of infrared modality and visible modality. Take the $n$-th local window $\hat{e}_j^v(n)$ and $\hat{e}_j^i(n) \in R^{M^2 \times 96}$ in the visible and infrared image embedding as an example, the corresponding query, key, and value matrices $Q$, $K$ and $V \in R^{M^2 \times 96/h \times h}$ are computed as

$$Q = \hat{e}_j^v(n)P_Q,\ K = \hat{e}_j^i(n)P_K,\ V = \hat{e}_j^i(n)P_V \quad (4)$$

where $h$ is the number of heads in multi-head attention, which is set as 3, $P_Q, P_K$ and $P_V$ are the projection matrices in the spatial feature alignment module shared across windows. Then the value $V$ generated from the infrared image feature is aligned to the visible feature as follows,

$$A = \text{SoftMax}(QK^T/\sqrt{d} + B)V \quad (5)$$

where $B$ is the learnable relative positional encoding, and $d = 96/h$ is the number of channels in each head. $A$ is the multi-head cross attention (MCA) output for the local window $\hat{e}_j^i(n)$ and is considered as the aligned embedding result from the infrared image to the visible image.

### 3.4. Multi-modality Video Compression

In practical applications like autonomous driving, the sensors usually capture multi-modality video information for the downstream analysis tasks. Since our approach is very flexible, we also extend the proposed framework for multi-modality video compression.

The overall pipeline is shown in Fig. 4, $\hat{x}_t^i$ and $x_t^v$ represent the reconstructed infrared frame and to be encoded visible frame at time step $t$, respectively. Here we re-implement the existing learning-based video compression method FVC as our baseline method [21]. FVC follows the hybrid coding framework and employs the deformable convolution to estimate the motion information for the subsequent motion compensation and residual coding.

In our multi-modality video compression approach, we first use FVC to compress the infrared video sequence. For each frame $x_t^v$ from the visible video sequence, in addition to the original motion compensation based on the previous reconstructed frame $\hat{x}_{t-1}^v$, we further produce the aligned feature $\bar{f}_t^{v,a}$ based on the reconstructed infrared image $\hat{x}_t^i$ at the same time step. Here, we use the channel-wise alignment module discussed in Section 3.2 to align the feature in the infrared modality to that in the visible modality. Then we use one convolutional layer to fuse the motion compensated feature $\bar{f}_t^{v,m}$ from the previous visible frame $\hat{x}_{t-1}^v$ and the feature $\bar{f}_t^{v,a}$ from the infrared frame $\hat{x}_t^i$, and generate the final predicted feature $\bar{f}_t^v$, which will be encoded by using the following residual compression module. Since the auto-encoder network is also used in learning-based video compression systems, like FVC, to compress residual and motion information, our spatial alignment module can be easily integrated into the existing framework with better compression performance. In our implementation, we will exploit the correlation between the intermediate features in the

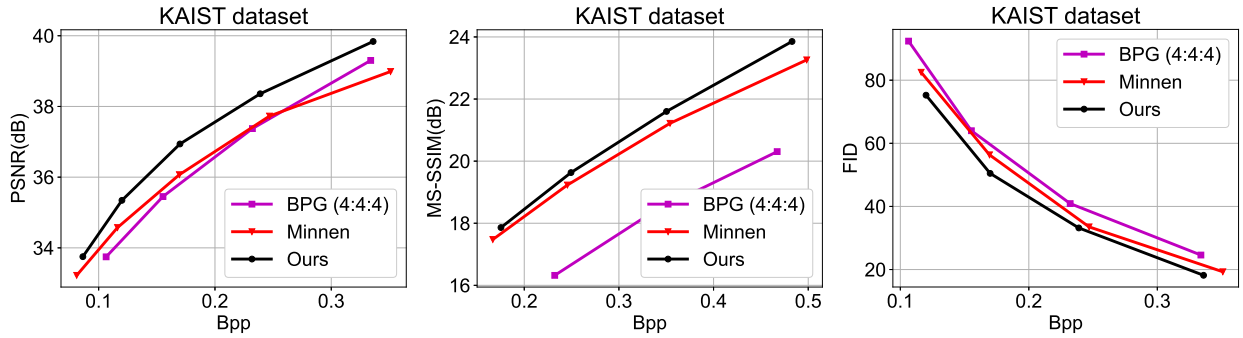


Figure 5. Visible image compression results from different approaches on the KAIST dataset in terms of PSNR, MS-SSIM and FID.

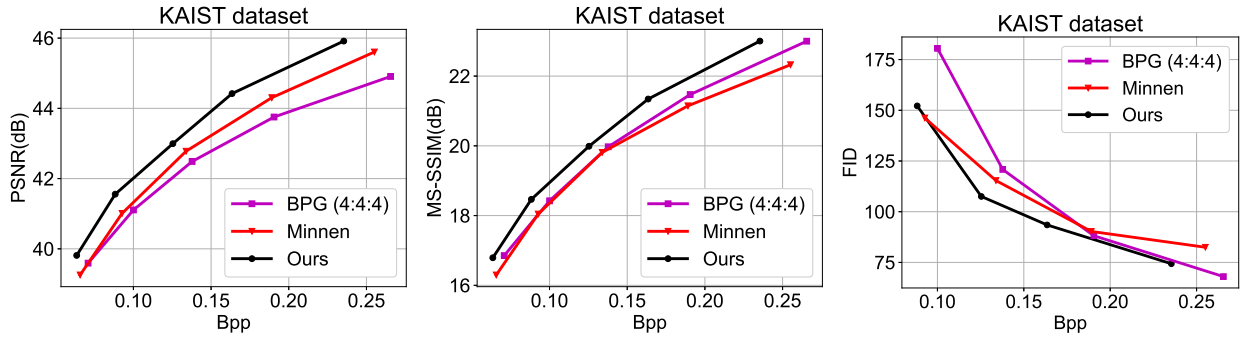


Figure 6. Infrared image compression results from different approaches on the KAIST dataset in terms of PSNR, MS-SSIM and FID.

residual decoder networks from different modalities. Due to the space limitation, we provide more implementation details for the multi-modality video compression in the supplementary materials. Although we use FVC [21] as an example to better introduce the proposed multi-modality compression framework, any other learning-based video compression approaches with both motion compensation and residual coding can also be integrated into our proposed framework.

# 4. Results

## 4.1. Experimental Setup

**FLIR Thermal Dataset [2]** It contains more than 10K pairs of 8-bit infrared (thermal) images and 24-bit visible images, including people, vehicles, bicycles, and other objects at both day and night scenes. The resolution of the infrared images is 640×512, while the corresponding resolution of visible images vary from 720×480 to 2048×1536. We resize each visible image to 1280×1024 in our experiments. The default FLIR training dataset is used as our training dataset, and 20 color-thermal pairs from the FLIR validation set are randomly selected as the testing dataset.

**KAIST Multispectral Pedestrian Dataset [22]** The dataset consists of 95K color-thermal pairs (640×480, 20Hz), containing 41 sequences from 12 sets. We employ 10 sets for multi-modality video training, while using the first 100 frames of each sequence in the other two sets(*set06* and *set10*) as our testing dataset. In addition, 18 color-thermal pairs are chosen from the KAIST dataset as another testing dataset for multi-modality image compression and the infrared images are resized to 320×240.

Table 1. The BDBR [7] results of our method and Minnen's approach when compared with BPG for the visible or infrared image compression on FLIR and KAIST datasets.

| Methods | FLIR | | KAIST | |
|---|---|---|---|---|
| | visible | infrared | visible | infrared |
| Minnen [30] | -22.342 | -14.960 | -3.624 | -8.751 |
| Ours | **-30.226** | **-21.621** | **-18.639** | **-21.289** |

**Evaluation Metrics** The bpp (bit per pixel) measures the average bits consumption in the compression procedure. In addition to the PSNR and MS-SSIM [43], we also use FID [20] metric, which is more consistent with human perception, to measure the distortion between the reconstructed image and the ground truth visible/infrared image.

**Implementation Details** When we use the infrared image as the reference to encode the visible image, we first train the infrared data compression network, and then we optimize the network for visible data compression by freezing the infrared image compression. These networks are implemented based on PyTorch with CUDA support and trained on a V100 GPU card. Specifically, for the multi-modality image compression, we set different $\lambda$ values ($\lambda$


Figure 7. Visual quality comparison for the visible image compression results from BPG [1], Minnen [30] and ours.

= 256, 512, 1024, 2048, 4096) and use the Adam optimizer [23] by setting the initial learning rate, $\beta_1$ and $\beta_2$ as 1e-4, 0.9, 0.999, respectively. The learning rate is reduced to 1e-5 after 1.8M steps when the loss becomes stable. The mini-batch size is set as 4. It takes about 8 days for the training stage. For the multi-modality image compression, we first train our model on the FLIR dataset and finetune the pretrain model on the KAIST training dataset to evaluate results for KAIST testing dataset. For the multi-modality video training, we first train the FVC model on the Vimeo-90k dataset by following its default setting [21] and finetune the model on the KAIST dataset for another 500K steps.

### 4.2. Experimental Results

**Multi-modality Image Compression** To demonstrate the effectiveness of our method, we compare our method with the traditional single-modality image compression method BPG [1] and the learned image compression approach proposed by Minnen *et al.* [30] on both FLIR and KAIST testing datasets. Besides, for fair comparison, both our model and the baseline method [30] are optimized using the same multi-modality data based on the MSE(*i.e.*, PSNR) metrics. The BDBR results are provided in Table 1.

Fig. 5 shows the rate-distortion curves from different compression approaches for visible image compression on the KAIST dataset. Compared with the separately optimized single-modality compression method [30], our approach using the infrared image as the reference can improve the compression performance by more than 0.7dB on the KAIST dataset. Besides, our approach also achieves much better compression performance than the traditional image compression method BPG. More results on the FLIR dataset are provided in the supplementary materials.

Similar results can also be observed for infrared image compression in Fig. 6, where we use the visible image as the reference. Our multi-modality compression approach has nearly 0.7dB improvement compared with the baseline method [30] without cross-modality reference.

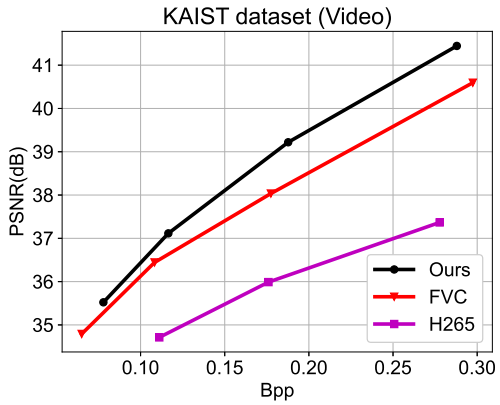


Figure 8. Experimental results from different video compression approaches on the KAIST dataset.

**Multi-modality Video Compression** In Fig. 8, we compare our method with the traditional single-modality codec H.265 [37] and the deep learning-based approach FVC [21]. The GoP size is set as 10 for the KAIST testing dataset. Similar to image compression, our model and the baseline model [21] are optimized on the MSE metrics on the same datasets. In addition, we follow the setting in [21] to produce the results of H.265. The experimental results show that when compared with the FVC and H.265, our approach has about 0.7dB and 3dB improvement by using the complementary information from infrared modality on visible sequences compression. It demonstrates that our proposed framework is very general and can be applied to the multi-modality image and video compression tasks.

### 4.3. Ablation Study

In Fig. 9, we provide the ablation study results on the KAIST dataset for the visible image compression. Here

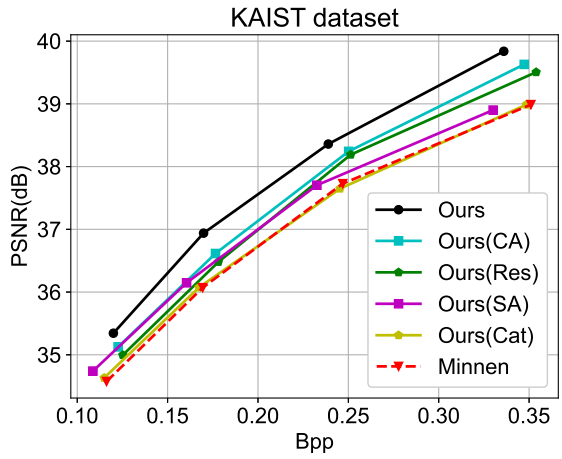


Figure 9. Ablation Study. *Ours(CA)* and *Ours(SA)* represent our models when only using the channel-wise alignment module and spatial-wise alignment module, respectively. *Ours* is our full model. *Ours(Res)* represents our model encoding residual information $f^v - \bar{f}^v$. *Ours(Cat)* represents simply concatenating $x^v$ and $x^i$ as the input for visible image compression.

*Ours(CA)* and *Ours(SA)* represent our proposed method using only channel-wise alignment module and spatial wise alignment module, respectively. And *Ours* denotes the full model in our implementation. It is observed that the channel-wise alignment module, *i.e*. *Ours(CA)*, can improve the compression performance by more than 0.4dB when compared with the baseline method. At the same time, it brings nearly 0.3dB gain by using the spatial alignment module (see *Ours(SA)*). Furthermore, our whole model has more than 0.7dB improvement over the baseline method by integrating both channel-wise and spatial-wise alignment modules. The experiments show that it is beneficial to exploit the complementary cross-modality information for the multi-modality data compression.

### 4.4. Model Analysis

**Element-wise Affine Transformation** In our implementation, we estimate the channel-wise affine transform coefficients. Here, we also provide the experimental results when using the element-wise affine transform. Specifically, we directly estimate the element-wise affine coefficients by removing the spatial-wise average pooling layer in Fig. 2(b). Considering the element-wise compression will consume a lot of bits, we further introduce another auto-encoder network to lossy compress these coefficients. Experimental results show that this new setting will deteriorate the compression performance by more than 6dB. The rate-distortion curve is provided in the supplementary material due to its much worse performance than the baseline method. One possible explanation is that the element-wise affine coefficients will consume much more bitrates, which leads to worse rate-distortion performance.

**Concatenation of the input multi-modality images.** We also provide a straightforward solution for multi-modality compression by concatenating the infrared and visible images as input to reconstruct the visible image. Experimental results show this straightforward solution (*Ours(Cat)*) has little improvement, which cannot reduce the cross-modality redundancy effectively.

**Residual Compression** In our framework, the output of the channel-wise feature alignment module is concatenated with the visible feature as the context information. In addition, we also try to compress the residual between aligned feature and visible feature, and found a performance drop of 0.14dB (see Fig. 9 *Ours(CA)* and *Ours(Res)*). In other words, although there is a certain amount of cross-modality redundancy between different modalities, it is difficult to compress the corresponding residual information.

**Qualitative Results** In Fig. 7, we also provide the qualitative results, and it is observed that our approach provides more visually plausible results. For example, the electric wire in the first-row image from our approach is much clearer when compared with that from BPG [1] or Minnen's approach [30].

**Joint Optimization** We also try to jointly optimize the infrared image compression and visible image compression in an end-to-end fashion. Experimental result shows that it only brings 0.02db improvement, and therefore we do not use it as we prefer a simple yet effective solution.

**Running Time and Complexity** The model parameters of our framework and the baseline method are 31M and 26M, respectively. We evaluate our framework with paired $1280 \times 1024$ visible image and $640 \times 512$ infrared image on a single V100 machine. The encoding speeds of our framework and the basic model are basically the same, and the decoding speeds are 340ms and 67ms, respectively.

## 5. Conclusions

In this work, we have proposed a multi-modality compression framework for visible and infrared image pairs. To exploit the complementary information, we introduce the channel-wise and spatial feature alignment modules. The experimental results on multiple benchmark datasets demonstrate the effectiveness of our multi-modality image and video compression approaches. In addition, our framework can also be extended for other multi-modality data that are close to each other, but may not work well for multi-modality data like images plus text descriptions and visual images plus point clouds as there is much less redundancy between different modalities. In the future, we will investigate new compression approaches for compressing more challenging multi-modality data.

**Acknowledgement** This work was supported by the National Natural Science Foundation of China under Grant 62102024.

## References


[1] F. bellard, bpg image format. http://bellard.org/bpg/. Accessed: 2018-10-30. 1, 2, 7, 8

[2] Flir thermal dataset. https://www.flir.com/oem/adas/adas-dataset-form/. Accessed: 2020-11-11. 6

[3] Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc V Gool. Soft-to-hard vector quantization for end-to-end learning compressible representations. In *NIPS*, pages 1141–1151, 2017. 1, 2

[4] Eirikur Agustsson, David Minnen, Nick Johnston, Johannes Balle, Sung Jin Hwang, and George Toderici. Scale-space flow for end-to-end optimized video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2020. 2

[5] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *5th International Conference on Learning Representations, ICLR*, 2017. 1, 2

[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. 1, 2

[7] Gisle Bjontegaard. Calculation of average psnr differences between rd-curves. *VCEG-M33*, 2001. 6

[8] Tahar Brahimi, Larbi Boubchir, Régis Fournier, and Amine Naït-Ali. An improved multimodal signal-image compression scheme with application to natural images and biomedical data. *Multimedia Tools and Applications*, 76(15):16783–16805, 2017. 2

[9] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 1, 2

[10] Ricardo Omar Chavez-Garcia and Olivier Aycard. Multiple sensor fusion and classification for moving object detection and tracking. *IEEE Transactions on Intelligent Transportation Systems*, 17(2):525–534, 2015. 1, 2

[11] Siqi Chen, Qiong Liu, and You Yang. Adaptive multimodality residual network for compression distorted multi-view depth video enhancement. *IEEE Access*, 8:97072–97081, 2020. 2

[12] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7939–7948, 2020. 2

[13] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1, 2

[14] Xin Deng, Wenzhe Yang, Ren Yang, Mai Xu, Enpeng Liu, Qianhan Feng, and Radu Timofte. Deep homography for efficient stereo image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1492–1501, 2021. 1, 2

[15] Abdelaziz Djelouah, Joaquim Campos, Simone Schaub-Meyer, and Christopher Schroers. Neural inter-frame compression for video coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6421–6429, 2019. 2

[16] Marek Fidali and Wojciech Jamrozik. Compression of high dynamic infrared image using auto aggregation algorithm. *Measurement Automation Monitoring*, 63, 2017. 2

[17] Adam Golinski, Reza Pourreza, Yang Yang, Guillaume Sautiere, and Taco S Cohen. Feedback recurrent autoencoder for video compression. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[18] AmirHossein Habibian, Ties van Rozendaal, Jakub M. Tomczak, and Taco Cohen. Video compression with rate-distortion autoencoders. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, pages 7032–7041. IEEE, 2019. 2

[19] Miska M Hannuksela, Ye Yan, Xuehui Huang, and Houqiang Li. Overview of the multiview high efficiency video coding (mv-hevc) standard. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 2154–2158. IEEE, 2015. 1, 2

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[21] Zhihao Hu, Guo Lu, and Dong Xu. Fvc: A new framework towards deep video compression in feature space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1502–1511, 2021. 2, 5, 6, 7

[22] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 6

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[24] Jin Li, Yao Fu, Guoning Li, and Zilong Liu. Remote sensing image compression in visible/near-infrared range using heterogeneous compressive sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(12):4932–4938, 2018. 2

[25] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun. Multi-task multi-sensor fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7345–7353, 2019. 1, 2

[26] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 641–656, 2018. 1, 2

[27] Jerry Liu, Shenlong Wang, and Raquel Urtasun. Dsic: Deep stereo image compression. In *Proceedings of the IEEE/CVF*

*International Conference on Computer Vision*, pages 3136–3145, 2019. 2

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 4

[29] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. DVC: An end-to-end deep video compression framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 11006–11015, 2019. 2

[30] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems*, pages 10771–10780, 2018. 1, 2, 3, 6, 7, 8

[31] Oren Rippel, Alexander G Anderson, Kedar Tatwawadi, Sanjay Nair, Craig Lytle, and Lubomir Bourdev. Elf-vc: Efficient learned flexible-rate video coding. *arXiv preprint arXiv:2104.14335*, 2021. 2

[32] Oren Rippel and Lubomir Bourdev. Real-time adaptive image compression. In *ICML*, 2017. 1, 2

[33] Oren Rippel, Sanjay Nair, Carissa Lew, Steve Branson, Alexander G. Anderson, and Lubomir D. Bourdev. Learned video compression. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, pages 3453–3462. IEEE, 2019. 2

[34] Heiko Schwarz, Detlev Marpe, and Thomas Wiegand. Overview of the scalable video coding extension of the h. 264/avc standard. *IEEE Transactions on circuits and systems for video technology*, 17(9):1103–1120, 2007. 1, 2

[35] Heiko Schwarz and Thomas Wiegand. Inter-view prediction of motion data in multiview video coding. In *2012 Picture Coding Symposium*, pages 101–104. IEEE, 2012. 2

[36] Athanassios Skodras, Charilaos Christopoulos, and Touradj Ebrahimi. The jpeg 2000 still image compression standard. *IEEE Signal Processing Magazine*, 18(5):36–58, 2001. 1, 2

[37] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, Thomas Wiegand, et al. Overview of the high efficiency video coding(hevc) standard. *TCSVT*, 22(12):1649–1668, 2012. 1, 2, 7

[38] George Toderici, Sean M. O'Malley, Sung Jin Hwang, Damien Vincent, David Minnen, Shumeet Baluja, Michele Covell, and Rahul Sukthankar. Variable rate image compression with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR*, 2016. 1, 2

[39] George Toderici, Damien Vincent, Nick Johnston, Sung Jin Hwang, David Minnen, Joel Shor, and Michele Covell. Full resolution image compression with recurrent neural networks. In *CVPR*, pages 5435–5443, 2017. 2

[40] Karthik Mahesh Varadarajan, Kai Zhou, and Markus Vincze. Rgb and depth intra-frame cross-compression for low bandwidth 3d video. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 955–958. IEEE, 2012. 2

[41] Anthony Vetro, Thomas Wiegand, and Gary J Sullivan. Overview of the stereo and multiview video coding extensions of the h. 264/mpeg-4 avc standard. *Proceedings of the IEEE*, 99(4):626–642, 2011. 2

[42] Gregory K Wallace. The jpeg still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1):xviii–xxxiv, 1992. 1, 2

[43] Zhou Wang, Eero Simoncelli, Alan Bovik, et al. Multi-scale structural similarity for image quality assessment. In *ASILOMAR CONFERENCE ON SIGNALS SYSTEMS AND COMPUTERS*, volume 2, pages 1398–1402. IEEE; 1998, 2003. 6

[44] Chao-Yuan Wu, Nayan Singhal, and Philipp Krahenbuhl. Video compression through image interpolation. In *ECCV*, September 2018. 2

[45] Yaojun Wu, Xin Li, Zhizheng Zhang, Xin Jin, and Zhibo Chen. Learned block-based hybrid image compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 2

[46] Ren Yang, Fabian Mentzer, Luc Van Gool, and Radu Timofte. Learning for video compression with hierarchical quality and recurrent enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6628–6637, 2020. 2

[47] Wenwei Zhang, Hui Zhou, Shuyang Sun, Zhe Wang, Jianping Shi, and Chen Change Loy. Robust multi-modality multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2365–2374, 2019. 1, 2

[48] Tong Zheng, Hirohisa Oda, Takayasu Moriya, Takaaki Sugino, Shota Nakamura, Masahiro Oda, Masaki Mori, Hirotsugu Takabatake, Hiroshi Natori, and Kensaku Mori. Multi-modality super-resolution loss for gan-based super-resolution of clinical ct images using micro ct image database. In *Medical Imaging 2020: Image Processing*, volume 11313, page 1131305. International Society for Optics and Photonics, 2020. 1, 2