remote sensing

MDPI

Article

MMCAN: Multi-Modal Cross-Attention Network for Free-Space Detection with Uncalibrated Hyperspectral Sensors

Feiyi Fang, Tao Zhou, Zhenbo Song and Jianfeng Lu *

School of Computer Science and Engineering, Nanjing University of Science and Technology, 200, Xiaolingwei, Nanjing 210094, China

* Correspondence: lujf@njust.edu.cn; Tel.: +86-025-8431-3997

Abstract: Free-space detection plays a pivotal role in autonomous vehicle applications, and its state-of-the-art algorithms are typically based on semantic segmentation of road areas. Recently, hyperspectral images have proven useful supplementary information in multi-modal segmentation for providing more texture details to the RGB representations, thus performing well in road segmentation tasks. Existing multi-modal segmentation methods assume that all the inputs are well-aligned, and then the problem is converted to fuse feature maps from different modalities. However, there exist cases where sensors cannot be well-calibrated. In this paper, we propose a novel network named multi-modal cross-attention network (MMCAN) for multi-modal free-space detection with uncalibrated hyperspectral sensors. We first introduce a cross-modality transformer using hyperspectral data to enhance RGB features, then aggregate these representations alternatively via multiple stages. This transformer promotes the spread and fusion of information between modalities that cannot be aligned at the pixel level. Furthermore, we propose a triplet gate fusion strategy, which can increase the proportion of RGB in the multiple spectral fusion processes while maintaining the specificity of each modality. The experimental results on a multi-spectral dataset demonstrate that our MMCAN model has achieved state-of-the-art performance. The method can be directly used on the pictures taken in the field without complex preprocessing. Our future goal is to adapt the algorithm to multi-object segmentation and generalize it to other multi-modal combinations.

Keywords: autonomous vehicles; semantic segmentation; multi-spectral data fusion; uncalibrated sensors

Citation: Fang, F.; Zhou, T.; Song, Z.; Lu, J. MMCAN: Multi-Modal Cross-Attention Network for Free-Space Detection with Uncalibrated Hyperspectral Sensors. *Remote Sens.* **2023**, *15*, 1142. https://doi.org/10.3390/rs15041142

Academic Editor: Claudio Piciarelli

Received: 22 December 2022
Revised: 16 February 2023
Accepted: 16 February 2023
Published: 20 February 2023

1. Introduction

As electric vehicles gradually replace traditional gasoline vehicles, the popularity of autonomous driving is also increasing year by year. People's awareness of autonomous vehicles has also shifted from science fiction to an everyday tool. Visual environment perception is the first link of autonomous driving, which helps autonomous vehicles to perceive and understand the surroundings [1]. Further, known as collision-free space detection, free-space detection is a fundamental component of visual environment perception. The approaches are generally semantic segmentation algorithms, which classify each pixel in an image into road or non-road classes. The segmentation results are then used by autonomous vehicles to navigate in complex environments and avoid obstacles.

In recent years, with the rapid development of computer technology, specifically the graphics processing unit (GPU), and the emergence of large-scale labeled data, the application of deep convolutional neural networks (DCNNs) has developed rapidly. It has become the mainstream method for free-space detection tasks. Thanks to the abundant data and accurate algorithms [2], it is convenient to train a segmentation DCNN. Even if the road is concealed by vehicles or under poor lighting conditions, these algorithms can provide a reliable result. Almost all standard road segmentation algorithms specifically support urban roads, which show either prominent boundary lines or clear texture demarcations in RGB images.

However, the segmentation method for visible-light images has limitations because of complex surface features in the wild or insufficient illumination at night. Such problems may be overcome by introducing hyperspectral imaging (HSI) or near-infrared (NIR) images. A spectral image with a resolution in the range of $10^{-2}\lambda$ is called a hyperspectral image [3]. Hyperspectral imaging is technology based on the continuous subdivision of narrow-band spectrums to simultaneously image the target area. It has become a mature technology that can capture detailed information for each pixel. Such a large amount of reflectance information about the underlying material can be helpful in accurate HSI segmentation. The hyperspectral images can help distinguish different substances, which is difficult in RGB images. Hence, HSI is widely used in various areas, including precision agriculture, military, surveillance, etc. [3,4]. Near-infrared is based on overtones and combinations of bond vibrations in molecules, a spectroscopic method that uses the near-infrared region of the electromagnetic spectrum. In NIR spectroscopy, light is absorbed in varying amounts by the object at particular frequencies corresponding to the combinations and overtones of vibrational frequencies of some bonds of the molecules in the object. Therefore, the bands seen in the NIR are typically extensive, leading to spectra that are more complex to interpret compared with RGB spectra. It generally penetrates deeper into an object's surface and can reveal the underlying material characteristics [5]. Thus, changes in intensity in the NIR image are due to material and illumination changes but not to color variations within the same material [6]. In the NIR image, the impact of the shadow on the road will be effectively suppressed, and the road area remains distinguishable in the dark. In order to achieve the segmentation task based on multiple spectral data, we believe that multi-modal machine learning (MMML) is a practical approach.

A modality refers to how something happens or is experienced. In this article, we regard modality as the data provided by sensors. Multi-modal perception aims to process and understand information from multi-source modalities. Learning from heterogeneous data brings the possibility of in-depth capturing correspondences. Examples are given in Figure 1 to show the advantage of multi-modal learning.

Figure 1. Example of real-world scenarios where current state-of-the-art single-modal approaches demonstrate misclassifications. The first row shows an issue of misclassifications caused by puddles that do not reflect the sky. The second row shows inconspicuous classes where roads and curbs are constructed of the same material.

Most existing multi-modal semantic segmentation methods are based on pixel-level aligned sensors, such as RGB and depth cameras, or multi-modal magnetic resonance imaging (MRI). This method provides a reasonable solution for unifying information from different modalities but is sensitive to the alignment of the input data. Unaligned multi-modal data will confuse the features learned by DCNNs, leading to false judgments, especially at low-dimensional layers. Today, public autonomous driving databases are dedicated to providing data for urban highway scenarios, and many free-space detection algorithms are customized based on such scenes. When these algorithms are transferred to

some particular scenes, such as rural or mountain roads, it is often difficult for the same effect to be achieved [7]. In order to achieve automatic driving tasks in these particular environments, we need to build an automatic driving collection platform and a database specific for rural or mountain roads. Different from the experiments performed on ready-made databases, such as KITTI [8] or Cityscapes [9], only uncalibrated data can be collected using a self-built experimental platform for multi-modal perception. In the data collection of autonomous driving, due to the different installation positions of multiple sensors, it is impossible to obtain completely aligned data from the source. The most common solution is calibrating the sensors and then performing the segmentation task using the aligned multi-modal information [10], while the mutual calibration of multiple sensors is complex work. In our experimental platform, three different spectral band sensors are included. Their field of view is adjusted to be as common as possible. The distortion of different camera lenses and different imaging principles makes pixel-level alignment of these three multi-spectral sensors impossible. Therefore, we explore a segmentation algorithm for uncalibrated multi-modal data to avoid extensive data calibration work.

To conquer this problem, we propose a cross-model transformer in a U-shape multi-modal semantic segmentation architecture, which fuses heterogeneous information and supports dynamic weighted feature fusion. Instead of alignment, we draw inspiration from representation and mapping methods that use uncalibrated sensors. Cross-attention [11] can be used to combine two embedding sequences regardless of their heterogeneity. In the cross-attention module, the similarity of the resulting points will reflect the semantic proximity between their corresponding original inputs. The attention mechanism for mixing two different embedding sequences in the transformer architecture requires that the two sequences have the same dimensionality but can be of different modalities. One of the sequences defines the output length as the query (Q), and the other sequence generates the key (K) and value (V). In our model, RGB is always input as Q, while hyperspectral sequences are always input as K and V. Since RGB road segmentation achieves satisfactory results for most scenes, we hope to make the RGB modality lead the multi-modal perception. Therefore, the features after embedding are then put into a gate fusion module [12]. After calculating the attention maps of the input features, a triplet gate is applied to obtain the adaptive RGB-guided fusion weights. Finally, the fused feature is sent into a segmentation decoder for the prediction result. Comprehensive experiments on the multi-spectral dataset HSI Road [13] show that our method provides excellent results in free-space detection tasks.

In this research, we directly exploit uncalibrated multi-modal data for the segmentation task. Our contributions in this paper are four-fold:

1. We propose a multi-modal free-space detection algorithm in an autonomous driving system with uncalibrated multi-spectral data.
2. We propose a cross-attention module that combines uncalibrated modalities. The attention mechanism extracts the relevant information of multi-modal data without pixel-wise alignment.
3. We design a multi-modal fusion architecture based on a triplet gate. In this structure, the participation of one primary modality is strengthened while the contributions of other modalities are maintained.
4. Experimental results on the HSI Road dataset demonstrate the effectiveness of the proposed multi-modal segmentation network compared with other existing approaches.

The rest of the paper is organized as follows: Section 2 summarizes the existing research on free-space detection and multi-modal feature fusion. Section 3 explains the proposed approach in detail. Section 4 provides details of the dataset and explains our experimental setup. Finally, Section 5 concludes the entire paper.

2. Related Work

We review some related work on free-space detection and multi-modal perception in the deployment of autonomous vehicle technology.

2.1. Free-Space Detection

Free-space detection is a binary pixel-level segmentation task. Popular single-modal semantic segmentation networks, such as FCN [14], SegNet [15], U-Net [16], PSPNet [17], DANet [18], etc., have achieved good performance for RGB free-space detection tasks. Today, state-of-the-art free-space detection networks usually use multi-modal data to assist RGB image segmentation and achieve excellent results, among which depth maps [19–26] or LiDAR point clouds [27–29] are the most commonly used modalities as they contain 3D information. SNE-RoadSeg+ [30] is the most representative one; it fuses RGB and dense disparity images and then obtains the segmentation result through a network with densely-connected skip connections, which achieves the state-of-the-art performance on the pioneering KITTI road [8] benchmark.

Although relatively rare, there are also some studies on multi-modal segmentation algorithms only using various 2D images. Shivakumar et al. [10] established an autonomous driving database containing RGB and thermal images, which is similar to the problem we face, but they have a different solution. They first performed calibration and then the segmentation process. Therefore, they also designed a two-stream segmentation architecture for the two modalities.

Due to its particular spectral range, NIR images often substitute RGB images for segmentation tasks under low-illumination conditions [31]. Before deep learning became popular, there were studies on combining NIR and RGB images for semantic segmentation [32,33]. In recent years, there have been studies on RGB+NIR for autonomous driving, using a dual-channel CNN model to perform semantic segmentation tasks for urban [34] and forest [35] scenes. Both of them used pixel-level aligned image data.

HSI images are mainly used for remote sensing tasks [36,37], but the algorithms for autonomous driving scenarios have not been well exploited. Huang et al. [38] applied HSI to semantic segmentation in cityscape scenes for the first time. They generated coarse labels with HSI images and utilized them to assist weakly supervised training with RGB images instead of fusing the two modalities.

2.2. Multi-Modal Feature Fusion

Multi-modal machine learning has been applied to various tasks, including speech synthesis [39,40], visual-audio recognition [41], sentiment analysis [42–44], image/video captioning [45–47], etc. As a part of multi-modal perception, most of the research on multi-modal segmentation [48–50] focuses on the feature fusion problem. Early works [19–21] on multi-modal learning concatenated calibrated images in different input channels to improve segmentation, which only required the training of a single model, making the training pipeline easy to construct. Other aspects [51,52] used single-modal decision values and fused them with a fusion mechanism.

Most commonly, multi-modal fusion is performed on latent features [22–24]. Dolz et al. [53] even proposed a densely connected network to connect and combine features from different layers of different modalities. This strategy of fusing pixels and features simultaneously allows the model to learn complex combined features between modalities freely. Chen et al. [54] introduced the method of feature gate fusion into multi-modal learning, which reduced the noise information in multi-modal data and allowed the incorporation of sufficiently complementary information to form discriminative representations for segmentation. However, these methods are all aimed at pixel-aligned feature maps.

Unfortunately, misalignment between multi-modal images is very common, but currently, no work can achieve multi-modal fusion from uncalibrated data for segmentation. In such conditions, Zhuang et al. [55] adopted a new label fusion algorithm for multi-modal images, which provided different levels of the structural information of images for multi-level local atlas ranking, utilized the information-theoretic measures to compute the similarity between modalities and performed the segmentation task after aligning the modalities. Chartsias et al. [56] corrected image misalignment with a Spatial Transformer Network and reconstructed the image to enable semi-supervised learning, thus bypass-

ing the problem of modal alignment. Joyce et al. [57] achieved MR image synthesis by encouraging the network to learn a modality invariant latent embedding during training to automatically correct misalignment in the input data, which has inspired us a lot. The study of modality embedding in this work inspired our approach to unaligned multi-modal data, but we believe that performing an image synthesis task is too complicated to guarantee high real-time ability in autonomous driving scenarios.

In the above research, although people are interested in using multi-spectral images and RGB images together for road detection tasks, the step of multi-modal image alignment is generally ignored since the images are preprocessed in the public dataset. However, in the actual autonomous driving scene, the installation method and imaging method of the sensors determine that multi-spectral images are difficult to align at the pixel level. We explore a model that could directly use unaligned multi-modal images so that it could be used on autonomous vehicles.

3. Method

To address the uncalibrated multi-modal free-space detection problem, we propose a novel network structure named multi-modal cross-attention network (MMCAN). To augment uncalibrated multi-spectral images with RGB data, we build a cross-modal encoder to enhance the modalities through multiple stages alternatively. The encoder utilizes a cross-attention module to project RGB features onto hyper-spectral features, which facilitates information propagation between modalities that are not aligned at the pixel level. We also applied a three-gate fusion strategy for multi-modal fusion to maintain the specificity of each modality.

In this section, we will first present the overall topology and training methods of the multi-modal free-space detection network. Secondly, we will introduce the proposed multi-modal cross-attention module. At last, we will describe the feature fusion details of the triplet gate.

3.1. Network Architecture

In our multi-modal free-space detection task, three kinds of data from different modalities as a group are input into the network, which are 3-channel RGB, 16-channel HSI, and 25-channel NIR images. Each group of multi-modal data corresponds to the same scene, but only the RGB image has ground truth. Therefore, our research focuses on extracting information from unaligned multi-modal images for the free-space detection task.

There are five research interests in multi-modal learning [58]: representation, translation, alignment, fusion, and co-learning. Multi-modal representation learning refers to summarizing the complementarity and eliminating the redundancy between multiple sensory modalities, including two representation methods. Joint representation means that the information of multiple sensory modalities is mapped to a unified multi-modal vector space. Coordinated representation means that each modality is mapped to its respective representation space, but the mapped vectors match certain relevance constraints. Transformation, also called mapping, is to transform the information of one modality into another. Alignment is to find the correspondence between elements of different modalities from the same instance. The alignment can be reflected in time and space. In image semantic segmentation tasks, the spatial alignment is reflected in each pixel of the picture corresponding to a semantic label. Multi-modal fusion is the combination of the information of multiple sensory modalities to perform a prediction, which is one of the earliest and most widely researched directions of multi-modal machine learning. According to the fusion level, multi-modal fusion has three categories: pixel level, feature level, and decision level, corresponding to the fusion of original data, the fusion of abstract features, and the fusion of decision results. Our studies usually focus on feature-level fusion. It includes early, middle, and late fusion approaches, which represent that the fusion occurs in the different stages of feature extraction. Co-learning is the transformation of knowledge

between different modalities. It can assist in the studies of multi-modal mapping, fusion, and alignment problems.

Multi-modal fusion is the key point in our research, which integrates information from different modalities into a stable multi-modal representation. The reason why multiple sensory modalities are needed to be integrated is that different modalities behave differently in the same scene, as there exist overlapping and complementarity, and even multiple different interactions between modalities. With well-processed multi-modal information, more abundant features can be obtained than single-modality, and the influence of redundant information will be reduced.

In this paper, we adopt the middle fusion strategy as the basis for our network design, which is to fuse the information at the feature level. Referring to the commonly used encoder-decoder structure, we design a separate encoder for each modality, which converts the input images into high-dimensional feature expressions, then integrates them before sending them into the segmentation decoder. As the selection of encoder, ResNet with residual block as a layer of feature extraction unit is our preferred structure. Its excellent feature extraction ability has been confirmed in numerous experiments. In order to fully preserve multi-scale features in segmentation, we design a U-shaped structure to connect the features to the decoder layer by layer. This is beneficial to the network's identification of segmentation edges.

We usually believe that in deep neural networks, the low-level features such as edges, contours, and colors contain visual information with less semantics but accurate location information, while the high-level features have rich semantic information, but their location is sketchy. Therefore, we place the feature fusion stage in the high-level layers of the encoder in order to prevent the network from learning pixel perturbations caused by misalignment in high-resolution images. Specifically, in the first two layers of the network, only RGB features are connected to the decoder through skip connections. While in the last three layers, RGB features are first used to aggregate with HSI and VIS features, respectively, then gate fused with the aggregated HSI and VIS features that are sent to the decoder in the end. The first two layers are more sensitive to details due to the smaller range of receptive fields; therefore, learning only RGB features with ground truth is sufficient for the network to predict the segmentation edges. For the last several layers with a larger range of receptive fields, the joint multi-modal features can effectively help the model learn high-dimensional semantic information and avoid misjudgments in road areas.

The overall structure of our MMCAN is depicted in Figure 2a. After the three types of spectra, images are passed through the ResNet [59] encoders. The feature maps of the HSI and NIR spectra are embedded in the RGB features, respectively, for heterogeneous information aggregation. The aggregated HSI and NIR feature maps will then be fused with the ResNet-encoded RGB feature in the last three layers through a triplet gate and sent to the U-shape decoder. The entire network is trained end-to-end, driven by cross-entropy loss defined on the segmentation benchmarks.

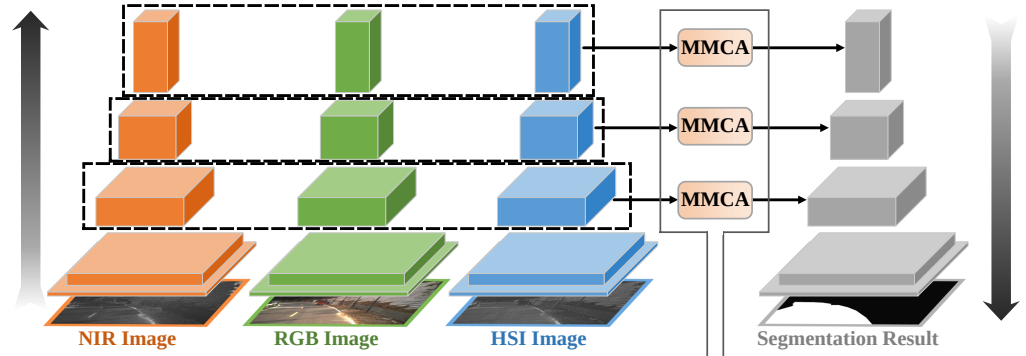
3.2. Multi-Modal Cross-Attention

Our multi-modal semantic segmentation needs to aggregate features from a group of uncalibrated multi-spectral images. The images in the same group correspond to different ground-truth. Learning features from a mismatching label confuses the representation learning system, resulting in convergent failure or wrong results. However, although each image in the same group is different in detail and size, the corresponding road scenes are almost the same. In the road segmentation task, our purpose is to minimize the misclassification of areas of the road rather than distinguish the edge details. Therefore, an effective cross-modality aggregation scheme should be able to extract effective segmentation information from this group of multi-modal data. We put forward a multi-modal cross-attention (MMCA) fusion to solve the problem. The framework of the proposed approach is shown in Figure 3. The fusion involves the RGB feature of one branch and HSI/NIR feature of the other branch. In order to fuse multi-modal features more efficiently and effectively,

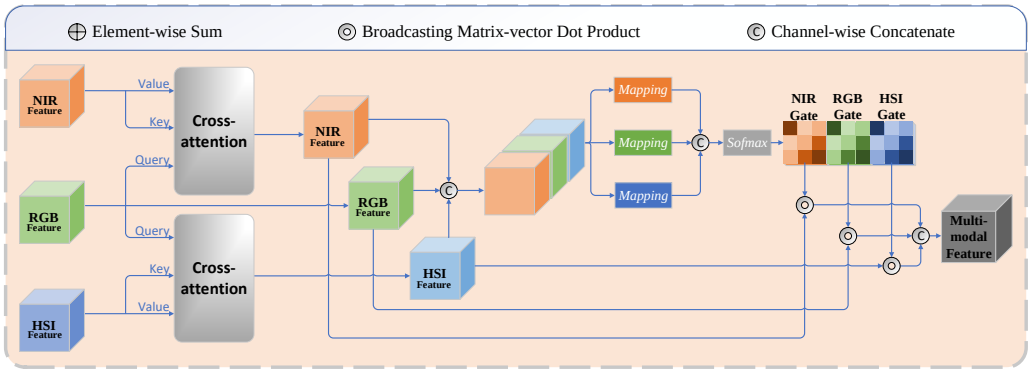
we utilize the RGB feature at each branch as an agent to exchange information among the multi-spectral feature from the other branch. The proposed operation can be precisely described in the Q-K-V language, namely matching a query from one modality with a set of key-value pairs from the other modality and thereby extracting the most critical cross-modality information. The MMCA operation consists of a set of queries $Q \in \mathbb{R}^{HW_1 \times d}$, and a set of keys $K \in \mathbb{R}^{HW_2 \times d}$ and values $V \in \mathbb{R}^{HW_2 \times d}$, where HW_1 is the pixel number of the query, HW_2 is the pixel number of key-value pairs, and d is the common dimensionality of all the input features. We calculate the dot products of the query with all keys, divide each by $\sqrt{d}$ and apply a softmax function to obtain the attention weights on the values. The MMCA operation can be mathematically expressed as:

$$Z = \text{MMCA}(Q, K, V) = softmax(QK^T/\sqrt{d})V, \quad (1)$$

where $Q \in \mathbb{R}^{HW_1 \times d}$ is the query, $K \in \mathbb{R}^{HW_2 \times d}$ is the key, and $V \in \mathbb{R}^{HW_2 \times d}$ is the value, and $Z \in \mathbb{R}^{HW_1 \times d}$ corresponds to the attended features of the queries.



(a) Overview of our network



(b) Details of the MMCA module

Figure 2. (**a**) Pipeline of the proposed segmentation network. An encoder-decoder architecture is employed. The input of the network is a group of RGB, HSI, and NIR images. They are processed by three encoders. In each group, the HSI and NIR features are weighted by the RGB feature separately, then fused by a triplet gate. The fusion result is propagated to a U-shape segmentation decoder for the final prediction. (**b**) Details of MMCA, including the implementation of two multi-modal cross-attention blocks and a triplet gate-fusion module.

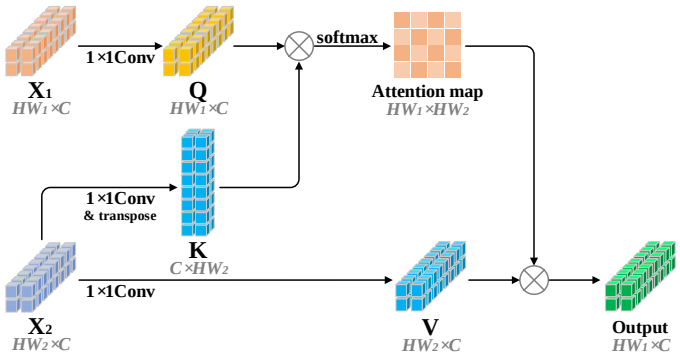


Figure 3. Cross-attention module for multi-modal features. The RGB feature map X_1 serves as a query token to interact with the patch tokens from multi-spectral features X_2. $\otimes$ denotes matrix multiplication.

In comparison with self-attention, which only pays attention to intra-modality, our proposed cross-modal attention allows the model to attend to diverse information from different modalities. Suppose $X_1 \in \mathbb{R}^{HW_1 \times C}$ and $X_2 \in \mathbb{R}^{HW_2 \times C}$ are from the feature maps of a specific stage of ResNet with dimension C. Q, K, and V are given as follows:

$$\{Q, K, V\} = \{X_1 \cdot W_q, X_2 \cdot W_k, X_2 \cdot W_v\}, \tag{2}$$

where $W_q, W_k, W_v \in \mathbb{R}^{C \times d}$ are learnable parameters of 1×1 convolutions. To prevent the model from becoming too large, we set $d = C/n$, where n is the reduction rate of the input dimension.

Implementation of multi-modal cross-attention. Figure 2b presents an example of the MMCA block with two cross-attention fusion streams. One stream is the aggregation of the HSI and RGB features, the other is for NIR and RGB. The two streams share the same structure but have independent training parameters.

Since the RGB image is the only annotated modality we have, Q comes from the RGB branch, and V, K come from the multi-spectral branches. This allows the RGB branches to participate in the overall position in the multi-spectral branch at a specific stage. As a result, it can selectively obtain more valuable information from possibly misaligned multi-spectral branches. The MMCA block can be added anywhere in CNNs because it can feed any value or key shape and ensure the same output shape as Q. This flexibility allows us to fuse richer layered features between uncalibrated modes. Thus, through the cross-attentional fusion operation, the latent features of the three modalities are aligned to $HW_1 \times C$.

3.3. Triplet Gate Fusion

The multi-spectral features are highly complementary, not only on the good side but also on the bad side. As the most widely used modality in free-space detection, RGB images provide rich and robust features for segmentation tasks. In fact, although multi-spectral images can provide more segmentation information than RGB images in some specific scenarios, segmentation using HSI or NIR modality alone cannot achieve the performance of an RGB modality-only network on the entire dataset. General fusion strategies, such as concatenation or summation, fuse the feature maps together without considering the disambiguation among modalities. For multi-modal learning, multi-source features of the same instance are mixed with each other, which may cause cross-modality ambiguity. In order to make full use of the complementarity of multi-modal information and filter the ambiguous features, we will selectively use them for fusion according to the presentation capabilities of different modalities. To this end, we design a triplet gate structure to measure the effectiveness of each modality and to fuse these features accordingly. The triplet gate is designed based on a concatenation-based fusion with a controlled information flow, which is visualized in Figure 2. The general idea of a gate fusion is that each feature map

$x_i \in \mathbb{R}^{C\times H\times W}$ is associated with a gate map $G_i \in [0,1]^{H\times W}$. A concatenation-based gate fusion can be defined as:

$$x = [x_i * G_i | i \in [1, M]], \tag{3}$$

where $M = 3$ is the number of feature maps. Specifically, we generate the triplet gate with the aggregated feature maps in the previous chapter, which are RGB $\in \mathbb{R}^{C\times H\times W}$ for RGB input, HSI $\in \mathbb{R}^{C\times H\times W}$ for HSI input, and NIR $\in \mathbb{R}^{C\times H\times W}$ for NIR input. The first step is to concatenate these three feature maps so as to collect their features in a specific dimension. The concatenated feature is then mapped to three different gate vectors with three convolutional layers F_{rgb}, F_{hsi} and F_{nir}:

$$V_{rgb} = F_{rgb}([\text{RGB}, \text{HSI}, \text{NIR}]) \in \mathbb{R}^{1\times H\times W}, \tag{4}$$

$$V_{hsi} = F_{hsi}([\text{RGB}, \text{HSI}, \text{NIR}]) \in \mathbb{R}^{1\times H\times W}, \tag{5}$$

$$V_{nir} = F_{nir}([\text{RGB}, \text{HSI}, \text{NIR}]) \in \mathbb{R}^{1\times H\times W}, \tag{6}$$

where V_{rgb}, V_{hsi}, and V_{nir} are three gate vectors of RGB, HSI, and NIR features, respectively. The three gate vectors are then concatenated to calculate the gate maps through a softmax function:

$$G_{rgb} = \frac{e^{V_{rgb}}}{e^{V_{rgb}} + e^{V_{hsi}} + e^{V_{nir}}} \in \mathbb{R}^{1\times H\times W}, \tag{7}$$

$$G_{hsi} = \frac{e^{V_{hsi}}}{e^{V_{rgb}} + e^{V_{hsi}} + e^{V_{nir}}} \in \mathbb{R}^{1\times H\times W}, \tag{8}$$

$$G_{nir} = \frac{e^{V_{nir}}}{e^{V_{rgb}} + e^{V_{hsi}} + e^{V_{nir}}} \in \mathbb{R}^{1\times H\times W}, \tag{9}$$

where the purpose is to normalize the gate maps G_{rgb}, G_{hsi}, and G_{nir} to meet the condition $G_{rgb} + G_{hsi} + G_{nir} = 1$, which represents the weights assigned to each position in the feature maps. The gate vectors are produced by a fully connected layer with a sigmoid function that adaptively controls the flow at the input. Therefore, the final fused feature X can be formulated as:

$$X = F_{map}([\text{RGB} * G_{rgb}, \text{HSI} * G_{hsi}, \text{NIR} * G_{nir}]), \tag{10}$$

where we join a 1×1 convolutional layer to map the feature vector X from $\mathbb{R}^{3C\times H\times W}$ to $\mathbb{R}^{C\times H\times W}$. Through this gate fusion module, the network has a robust feature retention mechanism to ensure that the decoders can learn complete information while eliminating the noise brought by the multi-modal data.

4. Experiments

Dataset. We evaluate our approach on the multi-spectral free-space detection dataset HSI Road [13]. It contains 3799 scenes with RGB, HSI, and NIR modalities, including 1811 rural scenes and 1988 urban scenes. All the modalities are respectively annotated, but we only use the RGB labels as the ground truth. The RGB modality used in the experiments is 3-channel 704×1280 pixel pictures, the HSI modality is 16-channel 256×480 pixel pictures, and the NIR modality is 25-channel 192×384 pixel pictures. Figure 4 shows the imaging characteristics of these spectra. Experiments are deployed on three sets, which are rural-only, urban-only, and all the datasets. Due to the small amount of data (less than 10,000), we set the ratio of the training set, test set, and validation set to 6:2:2. Therefore, for each experiment, we randomly use 60% data as the training set, 20% as the testing set, and the remaining 20% as the validation set.

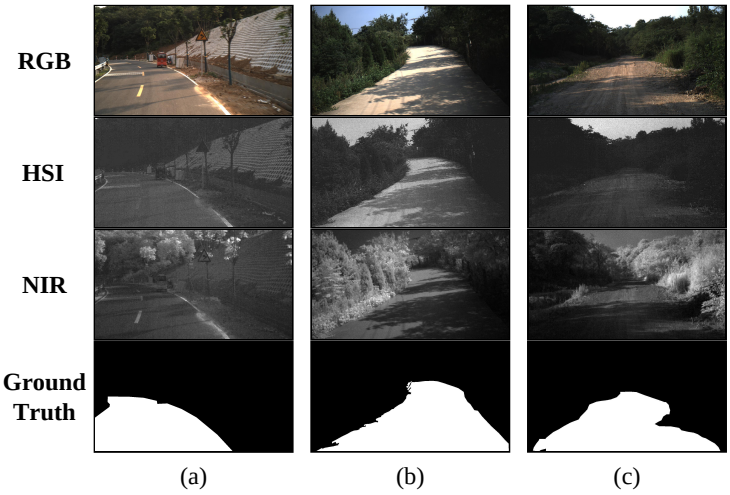


Figure 4. Example of multi-spectral images in HSI Road dataset [13]. (**a**–**c**) show three different scenes, and each scene includes three images that are from RGB, HSI, and NIR, respectively (from up to down). The images on the bottom represent the ground truth, which is annotated according to the RGB spectrum.

Implementation Details. Our network is implemented by Pytorch and trained on NVIDIA Tesla V100 (Nvidia, CA, USA) platform using CUDA10.0. Our batch size is set to 6, the initial learning rate is set to 1×10^{-4}, and the Adam solver is used to optimize the network. We train the network over 100 epochs and decay the learning rate linearly at a rate of 0.99.

Evaluation Metrics. Free-space detection is a two-class segmentation problem. Following recent methods, we employ two metrics to evaluate the performance of our networks, such as pixel accuracy and mIoU. The metrics are computed as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{11}$$

$$\text{mIoU} = \frac{2TP}{TP + FP + FN} + \frac{2TN}{TN + FN + FP}, \tag{12}$$

where TP, TN, FP, and FN represent the number of true positive, true negative, false positive, and false negative pixels. The results from these formulas are dimensionless. The Accuracy will show the ratio of correct predicting pixels, and the mIoU will show the ratio of intersection and union of ground truth and predicted results.

4.1. Experimental Results

In our experiments, we compare our MMCAN with SOTA semantic segmentation approaches. We use the dataset to train ten DCNNs, including five single-modal networks and four multi-modal networks. The approaches are tested under three settings: (a) training with urban scenes, (b) training with rural scenes, and (c) training with mixture scenes. The single-modal experiments are conducted with RGB images only. The multi-modal experiments are conducted with two fusion strategies: early fusion and middle fusion.

For single-modal experiments, we implemented two baseline segmentation approaches, i.e., U-Net [16] and DeepLab-v3 [60], and deployed three SOTA methods, i.e., DANet [18], HRNet [61] and Self-Regulation [62]. The backbone of HRNet is set to HRNetV2-W48, and the others are ResNet-50. In the task of multi-modal learning, early fusion methods indicate a U-Net with a concatenation of images as input, middle fusion methods include HAFB [50] and a multi-encoder U-Net baseline called MU-Net [63], which consists of three independent ResNet-50 encoders for the three modalities and the feature maps of each layer are concatenated to fuse as the skip connections of a U-Net decoder. To compare

the performances between our proposed method and other SOTA DCNNs, we train our MMCAN with the same setup as for the multi-modal networks.

We evaluate the performance of our proposed MMCAN qualitatively and quantitatively. The comparisons of accuracy and mIoU scores on the validation set are shown in Table 1. It can be observed that the results show that the score in rural scenes is lower than that in urban scenes, while the score is between them under the entire dataset. The scores of SOTA multi-modal learning are similar to that of the SOTA single-modal network in the urban scene and increase by 0.5–5% in the rural scene, which indicates that the multi-modal data can indeed make up for the deficiencies of the RGB modality. Our proposed MMCAN outperforms the RGB-based single-modal methods and also multi-modal methods designed for aligned images for both urban and rural scenarios, with a score gain of 1.2–4.5%.

Table 1. Performance on the HSI Road validation set, divided into three conditions: urban scenes, rural scenes, and all scenes.

Methods		Accuracy			mIoU		
		Urban	**Rural**	**All**	**Urban**	**Rural**	**All**
Single-Modal (RGB)	U-Net [16]	95.88	93.82	94.39	92.87	90.25	92.12
	Deeplab-V3 [17]	95.86	93.87	94.33	93.18	90.74	92.90
	DANet [18]	97.03	94.86	95.52	94.96	91.89	93.33
	HRNet [61]	97.49	94.74	96.42	94.46	91.77	93.20
	Self-Regulation [62]	98.05	96.00	97.32	95.40	92.48	94.68
Multi-Modal	U-Net [16]	97.08	96.38	97.14	94.16	93.04	94.18
	MU-Net [63]	97.88	95.82	97.39	95.70	92.48	94.88
	HAFB [50]	98.10	95.95	97.32	96.64	93.98	95.29
	MMCAN	**98.68**	**97.78**	**98.29**	**97.36**	**95.35**	**96.41**

Examples of the experimental results on the HSI Road dataset are shown in Figure 5. We can clearly observe that single-modal methods with RGB images as inputs can usually generate pretty accurate segmentation results, but it also suffers from occasional misclassification due to poor shadow and lighting conditions. Early fusion and intermediate fusion strategies using aligned data can effectively improve performance, recovering rough road shapes but with inaccurate segmentation boundaries. Our approach takes into account the above two points, not only presenting more accurate free-space estimations but also ensuring the details of the boundaries.

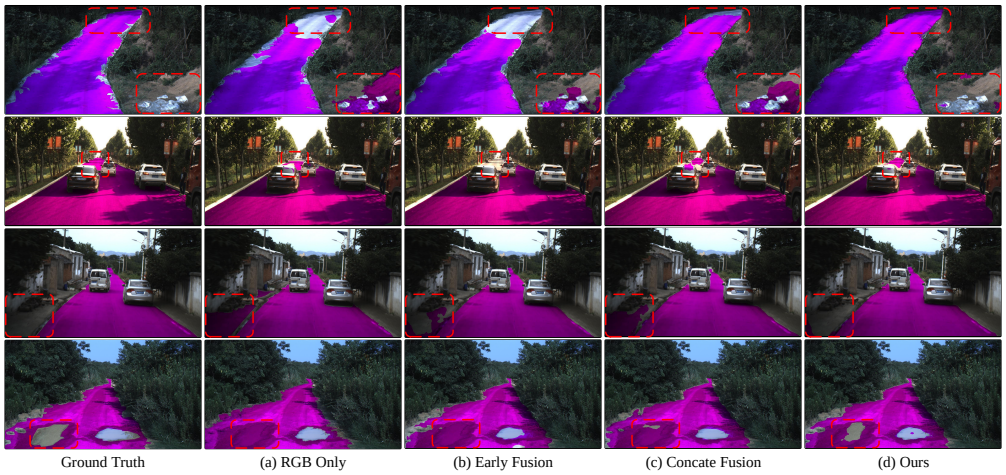


Figure 5. Examples on the HSI road dataset, where (**a**–**d**) show the segmentation results obtained by single-modal HRNet, early fusion U-Net, concatenate fusion MU-Net, and our proposed MMCAN, respectively.

The experimental results show that our method has three advantages. First, in the urban environment, the method is as good as the SOTA RGB single-modal method, with slightly higher accuracy; by 0.63%. Secondly, in the rural environment, the method has obvious advantages compared with the RGB single-mode method, with a 1.78% higher score. This is because the rural environment is unstructured; thus there are many features that cannot be perceived by RGB cameras, and the task can only be completed with the supplement of multi-spectral information. Thirdly, compared with other multi-modal methods, the method uses multi-modal cross-attention to solve the problem of data alignment and can directly process unaligned multi-spectral data. However, the method also has some disadvantages. It is insufficient in the accuracy of segmentation edges, and, at the same time, it has defects in recognition of small targets, which needs further research and exploration in the future.

4.2. Ablation Study

To validate the effectiveness of every component in the proposed MMCAN, we performed ablation experiments on the HSI Road dataset.

First, we investigate the impact of concatenation fusion and our proposed triplet gate fusion by replacing the gate fusion blocks with concatenation operators. As shown in Table 2, the gate fusion strategy significantly outperforms the simple concatenation fusion strategy for multi-modal free-space detection, the performance is increased by 0.61 points in urban scenes and 1.32 points in the whole dataset, which can be attributed to the fact that the gate reduces noise in the modalities, and useful information is emphasized as a result.

Table 2. Ablation study on fusion strategies.

Methods	Accuracy			mIoU		
	Urban	**Rural**	**All**	**Urban**	**Rural**	**All**
MMCAN + concatenation	98.19	**97.83**	97.52	96.75	**95.56**	95.09
MMCAN + triplet gate	**98.68**	97.78	**98.29**	**97.36**	95.35	**96.41**

Then, we remove the inputs from MMCAN to evaluate its performance on single-modal vision data. We conduct three experiments: (a) training with RGB images, (b) training with HSI images, (c) training with NIR images, (d) training with RGB + HSI modalities, and (e) training with RGB + NIR modalities. From Table 3, we can observe that our choice outperforms the single-modal architecture concerning different modalities of training data, proving that the data fusion via a three-encoder architecture can benefit from free-space detection. It should be noted that although in the single-modal condition, our approach cannot provide competitive results, the network still achieves sufficiently reliable segmentation.

Table 3. Ablation study on different modalities.

Methods	Modalities	Accuracy			mIoU		
		Urban	**Rural**	**All**	**Urban**	**Rural**	**All**
U-Net	RGB	95.88	93.82	94.39	92.87	90.25	92.12
	HSI	84.37	85.76	89.00	86.58	82.22	88.59
	NIR	87.27	89.94	92.53	84.64	85.52	87.40
	RGB + HSI	95.48	92.79	96.33	91.25	88.73	92.94
	RGB + NIR	94.73	94.17	95.52	91.21	89.53	92.12
	RGB + NIR + HSI	**97.08**	**96.38**	**97.14**	**94.16**	**93.04**	**94.18**
MMCAN	RGB	95.08	94.38	95.14	92.16	90.04	92.18
	HSI	85.85	82.45	91.23	84.42	84.26	86.17
	NIR	87.41	81.88	88.57	82.32	85.17	87.56
	RGB + HSI	98.48	96.45	97.12	94.48	93.20	94.85
	RGB + NIR	98.24	91.72	96.57	95.67	92.50	93.89
	RGB + NIR + HSI	**98.68**	**97.78**	**98.29**	**97.36**	**95.35**	**96.41**

To further validate the effectiveness of our choice, we add the MMCA module to the low-dimensional layers of the network. Table 4 verifies the superiority of deploying MMCA modules in high-dimensional layers, which helps to alleviate feature confusion to generate accurate free-space detection results.

Table 4. Ablation study on MMCA module.

Layers	mIoU		
	Urban	**Rural**	**All**
1 + 2 + 3 + 4 + 5	95.78	92.73	94.91
2 + 3 + 4 + 5	96.44	94.24	95.27
3 + 4 + 5 (ours)	**97.36**	**95.35**	**96.41**
4 + 5	97.18	94.92	96.03
5	96.65	93.71	94.97

5. Conclusions

In this paper, we have presented a cross-modality embedding aggregation network that can be used for free-space detection tasks on uncalibrated multi-spectral images, which is a combination of sensors deployed on autonomous vehicles. Unlike existing multi-modal segmentation methods, this network does not rely on pixel-wise aligned images; therefore, a lot of preprocessing work, such as calibration and labeling, can be reduced. The network is able to correct the erroneous results of RGB single-modal segmentation in specific scenarios. Meanwhile, the joint triplet gate fusion can eliminate the ambiguous information of multi-modal data. The experimental results on HSI, NIR, and RGB tri-modal dataset show that our model not only has a significant improvement in rural and mountain scenes but also achieves SOTA in multi-scene training. The model provides a solution for multi-modal perception in autonomous driving without data preprocessing, which greatly alleviates the computational cost. There are still deficiencies in our work. The model predicts segmentation edges imprecisely and performs poorly in the detection of tiny objects. Our future work focuses on two points. The first one is to extend the algorithm to other autonomous driving tasks, such as multi-target segmentation, prediction, and 3D segmentation. The other is to explore solutions to misaligned modalities in other multi-modal vision problems.

Author Contributions: Conceptualization, F.F. and T.Z.; methodology, F.F.; software, F.F.; validation, F.F.; formal analysis, T.Z.; investigation, F.F.; resources, J.L.; data curation, J.L.; writing—original draft preparation, F.F.; writing—review and editing, F.F.; visualization, Z.S.; supervision, J.L.; project administration, T.Z.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was sponsored by Jiangsu Funding Program for Excellent Postdoctoral Talent 2022ZB268.

Data Availability Statement: The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DCNNs Deep Convolutional Neural Networks
MMML Multi-Modal Machine Learning

References

1. Zhu, H.; Yuen, K.-V.; Mihaylova, L.; Leung, H. Overview of environment perception for intelligent vehicles. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 2584–2601. [CrossRef]
2. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
3. Adão, T.; Hruška, J.; Pádua, L.; Bessa, J.; Peres, E.; Morais, R.; Sousa, J.J. Hyperspectral imaging: A review on uav-based sensors, data processing and applications for agriculture and forestry. *Remote Sens.* **2017**, *9*, 1110. [CrossRef]
4. Dou, H.-X.; Lu, X.-S.; Wang, C.; Shen, H.-Z.; Zhuo, Y.-W.; Deng, L.-J. Patchmask: A data augmentation strategy with gaussian noise in hyperspectral images. *Remote Sens.* **2022**, *14*, 6308. [CrossRef]
5. Timmer, B.; Reshitnyk, L.Y.; Hessing-Lewis, M.; Juanes, F.; Costa, M. Comparing the use of red-edge and near-infrared wavelength ranges for detecting submerged kelp canopy. *Remote Sens.* **2022**, *14*, 2241. [CrossRef]
6. Fedorov, S.; Molkov, A.; Kalinskaya, D. Aerosol optical properties above productive waters of gorky reservoir for atmospheric correction of sentinel-3/olci images. *Remote Sens. 2022*, *14*, 6130. [CrossRef]
7. Zhang, X.; Li, Z.; Gong, Y.; Jin, D.; Li, J.; Wang, L.; Zhu, Y.; Liu, H. Openmpd: An open multimodal perception dataset for autonomous driving. *IEEE Trans. Veh. Technol.* **2022**, *71*, 2437–2447. [CrossRef]
8. Fritsch, J.; Kuehnl, T.; Geiger, A. A new performance measure and evaluation benchmark for road detection algorithms. In Proceedings of the International Conference on Intelligent Transportation Systems, The Hague, The Netherlands, 6–9 October 2013.
9. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
10. Shivakumar, S.S.; Rodrigues, N.; Zhou, A.; Miller, I.D.; Kumar, V.; Taylor, C.J. Pst900: Rgb-thermal calibration, dataset and segmentation network. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation, Paris, France, 31 May–31 August 2020; pp. 9441–9447.
11. Wei, X.; Zhang, T.; Li, Y.; Zhang, Y.; Wu, F. Multi-modality cross attention network for image and sentence matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10941–10950.
12. Wang, H.; Wang, Y.; Zhang, Q.; Xiang, S.; Pan, C. Gated convolutional neural network for semantic segmentation in high-resolution images. *Remote Sens.* **2017**, *9*, 446. [CrossRef]
13. Lu, J.; Liu, H.; Yao, Y.; Tao, S.; Tang, Z.; Lu, J. Hsi road: A hyper spectral image dataset for road segmentation. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo, London, UK, 6–10 July 2020; pp. 1–6.
14. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
15. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
16. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
17. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
18. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
19. Couprie, C.; Farabet, C.; Najman, L.; LeCun, Y. Indoor semantic segmentation using depth information. *arXiv* **2013**, arXiv:1301.3572.
20. Hazirbas, C.; Ma, L.; Domokos, C.; Cremers, D. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 213–228.
21. Eigen, D.; Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE international Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2650–2658.
22. Lin, D.; Fidler, S.; Urtasun, R. Holistic scene understanding for 3d object detection with rgbd cameras. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1417–1424.
23. Li, Z.; Gan, Y.; Liang, X.; Yu, Y.; Cheng, H.; Lin, L. Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 541–557.
24. Park, S.-J.; Hong, K.-S.; Lee, S. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4980–4989.
25. Wang, H.; Fan, R.; Sun, Y.; Liu, M. Applying surface normal information in drivable area and road anomaly detection for ground mobile robots. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems, Las Vegas, NV, USA, 25–29 October 2020; pp. 2706–2711.
26. Gu, S.; Zhang, Y.; Tang, J.; Yang, J.; Alvarez, J.M.; Kong, H. Integrating dense lidar-camera road detection maps by a multi-modal crf model. *IEEE Trans. Veh. Technol.* **2019**, *68*, 11635–11645. [CrossRef]

27. Chen, Z.; Zhang, J.; Tao, D. Progressive lidar adaptation for road detection. *IEEE/CAA J. Autom. Sin.* **2019**, *6*, 693–702. [CrossRef]
28. Caltagirone, L.; Bellone, M.; Svensson, L.; Wahde, M. Lidar–camera fusion for road detection using fully convolutional neural networks. *Robot. Auton. Syst.* **2019**, *111*, 125–131. [CrossRef]
29. Gu, S.; Yang, J.; Kong, H. A cascaded lidar-camera fusion network for road detection. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation, Xi'an, China, 30 May–5 June 2021; pp. 13308–13314.
30. Wang, H.; Fan, R.; Cai, P.; Liu, M. Sne-roadseg+: Rethinking depth-normal translation and deep supervision for freespace detection. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems, Prague, Czech Republic, 27 September–1 October 2021; pp. 1140–1145.
31. Qi, L.; Hu, Z.; Zhou, X.; Ni, X.; Chen, F. Multi-sensor fusion of sdgsat-1 thermal infrared and multispectral images. *Remote Sens.* **2022**, *14*, 6159. [CrossRef]
32. Salamati, N.; Larlus, D.; Csurka, G.; Süsstrunk, S. Semantic image segmentation using visible and near-infrared channels. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 461–471.
33. Salamati, N.; Larlus, D.; Csurka, G.; Süsstrunk, S. Incorporating near-infrared information into semantic image segmentation. *arXiv* **2014**, arXiv:1406.6147.
34. Choe, G.; Kim, S.-H.; Im, S.; Lee, J.-Y.; Narasimhan, S.G.; Kweon, I.S. Ranus: Rgb and nir urban scene dataset for deep scene parsing. *IEEE Robot. Autom. Lett.* **2018**, *3*, 1808–1815. [CrossRef]
35. Valada, A.; Oliveira, G.L.; Brox, T.; Burgard, W. Deep multispectral semantic scene understanding of forested environments using multimodal fusion. In Proceedings of the International Symposium on Experimental Robotics, Nagasaki, Japan, 3–8 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 465–477.
36. Sun, L.; Song, X.; Guo, H.; Zhao, G.; Wang, J. Patch-wise semantic segmentation for hyperspectral images via a cubic capsule network with emap features. *Remote Sens.* **2021**, *13*, 3497. [CrossRef]
37. Shen, X.; Weng, L.; Xia, M.; Lin, H. Multi-scale feature aggregation network for semantic segmentation of land cover. *Remote Sens.* **2022**, *14*, 6156. [CrossRef]
38. Huang, Y.; Shen, Q.; Fu, Y.; You, S. Weakly-supervised semantic segmentation in cityscape via hyperspectral image. In Proceedings of the IEEE International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 1117–1126.
39. Luong, H.-T.; Yamagishi, J. Multimodal speech synthesis architecture for unsupervised speaker adaptation. *arXiv* **2018**, arXiv:1808.06288.
40. Ma, S.; McDuff, D.; Song, Y. Unpaired image-to-speech synthesis with multimodal information bottleneck. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October– 2 November 2019; pp. 7598–7607.
41. Hou, J.-C.; Wang, S.-S.; Lai, Y.-H.; Tsao, Y.; Chang, H.-W.; Wang, H.-M. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Trans. Emerg. Top. Comput.* **2018**, *2*, 117–128. [CrossRef]
42. Poria, S.; Cambria, E.; Howard, N.; Huang, G.-B.; Hussain, A. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* **2016**, *174*, 50–59. [CrossRef]
43. Soleymani, M.; Garcia, D.; Jou, B.; Schuller, B.; Chang, S.-F.; Pantic, M. A survey of multimodal sentiment analysis. *Image Vis. Comput.* **2017**, *65*, 3–14. [CrossRef]
44. Deng, D.; Zhou, Y.; Pi, J.; Shi, B.E. Multimodal utterance-level affect analysis using visual, audio and text features. *arXiv* **2018**, arXiv:1805.00625.
45. Yu, J.; Li, J.; Yu, Z.; Huang, Q. Multimodal transformer with multi-view visual representation for image captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 4467–4480. [CrossRef]
46. Song, J.; Guo, Y.; Gao, L.; Li, X.; Hanjalic, A.; Shen, H.T. From deterministic to generative: Multimodal stochastic rnns for video captioning. *IEEE Trans. Neural Netw. Learn.* **2018**, *30*, 3047–3058. [CrossRef]
47. Xu, J.; Yao, T.; Zhang, Y.; Mei, T. Learning multimodal attention lstm networks for video captioning. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 537–545.
48. Feng, D.; Haase-Schütz, C.; Rosenbaum, L.; Hertlein, H.; Glaeser, C.; Timm, F.; Wiesbeck, W.; Dietmayer, K. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 1341–1360. [CrossRef]
49. Valada, A.; Mohan, R.; Burgard, W. Self-supervised model adaptation for multimodal semantic segmentation. *Int. J. Comput.* **2020**, *128*, 1239–1285. [CrossRef]
50. Fang, F.; Yao, Y.; Zhou, T.; Xie, G.; Lu, J. Self-supervised multi-modal hybrid fusion network for brain tumor segmentation. *IEEE J. Biomed. Health Inform.* **2021**, *26*, 5310–5320. [CrossRef]
51. Aasen, H.; Honkavaara, E.; Lucieer, A.; Zarco-Tejada, P.J. Quantitative remote sensing at ultra-high resolution with uav spectroscopy: A review of sensor technology, measurement procedures, and data correction workflows. *Remote Sens.* **2018**, *10*, 1091. [CrossRef]
52. Mu, C.; Dong, Z.; Liu, Y. A two-branch convolutional neural network based on multi-spectral entropy rate superpixel segmentation for hyperspectral image classification. *Remote Sens.* **2022**, *14*, 1569. [CrossRef]
53. Dolz, J.; Gopinath, K.; Yuan, J.; Lombaert, H.; Desrosiers, C.; Ayed, I.B. Hyperdense-net: A hyper-densely connected cnn for multi-modal image segmentation. *IEEE Trans. Med. Imaging* **2018**, *38*, 1116–1126. [CrossRef] [PubMed]

54. Chen, X.; Lin, K.-Y.; Wang, J.; Wu, W.; Qian, C.; Li, H.; Zeng, G. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 561–577.
55. Zhuang, X.; Shen, J. Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. *Med. Image Anal.* **2016**, *31*, 77–87. [CrossRef] [PubMed]
56. Chartsias, A.; Papanastasiou, G.; Wang, C.; Semple, S.; Newby, D.E.; Dharmakumar, R.; Tsaftaris, S.A. Disentangle, align and fuse for multimodal and semi-supervised image segmentation. *IEEE Trans. Med. Imaging* **2020**, *40*, 781–792. [CrossRef] [PubMed]
57. Joyce, T.; Chartsias, A.; Tsaftaris, S.A. Robust multi-modal mr image synthesis. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Quebec City, QC, Canada, 10–14 September 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 347–355.
58. Baltrušaitis, T.; Ahuja, C.; Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443. [CrossRef] [PubMed]
59. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
60. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
61. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach.* **2020**, *43*, 3349–3364. [CrossRef]
62. Zhang, D.; Zhang, H.; Tang, J.; Hua, X.-S.; Sun, Q. Self-regulation for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 6953–6963.
63. Dolz, J.; Desrosiers, C.; Ayed, I.B. Ivd-net: Intervertebral disc localization and segmentation in mri with a multi-modal unet. In Proceedings of the International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging, Granada, Spain, 16 September 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 130–143.