Robust Deep Multi-modal Learning Based on Gated Information Fusion Network

Jaekyum Kim[1], Junho Koh[1], Yecheol Kim[1], Jaehyung Choi[2], Youngbae Hwang[3], and Jun Won Choi[1(✉)]

[1] Hanyang University, Seoul, Korea
{jkkim,jhkoh,yckim}@spa.hanyang.ac.kr, junwchoi@hanyang.ac.kr
[2] Phantom AI Inc., Burlingame, CA, USA
jaehyung@phantom.ai
[3] Korea Electronics Technology Institute (KETI), Seongnam-si, Korea
ybhwang@keti.re.kr

Abstract. The goal of multi-modal learning is to use complementary information on the relevant task provided by the multiple modalities to achieve reliable and robust performance. Recently, deep learning has led significant improvement in multi-modal learning by allowing for fusing high level features obtained at intermediate layers of the deep neural network. This paper addresses a problem of designing robust deep multi-modal learning architecture in the presence of the modalities degraded in quality. We introduce deep fusion architecture for object detection which processes each modality using the separate convolutional neural network (CNN) and constructs the joint feature maps by combining the intermediate features obtained by the CNNs. In order to facilitate the robustness to the degraded modalities, we employ the gated information fusion (GIF) network which weights the contribution from each modality according to the input feature maps to be fused. The combining weights are determined by applying the convolutional layers followed by the sigmoid function to the concatenated intermediate feature maps. The whole network including the CNN backbone and GIF is trained in an end-to-end fashion. Our experiments show that the proposed GIF network offers the additional architectural flexibility to achieve the robust performance in handling some degraded modalities.

Keywords: Object detection · Multi-modal fusion · Sensor fusion · Gated information fusion

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government(MSIT) (2016-0-00564, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding).

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-20870-7_6) contains supplementary material, which is available to authorized users.

C. V. Jawahar et al. (Eds.): ACCV 2018, LNCS 11364, pp. 90–106, 2019.
https://doi.org/10.1007/978-3-030-20870-7_6

1 Introduction

Multi-modal learning refers to a machine learning problem aiming to improve learning performance using the experience acquired from the different types of data sources. Basically, such multi-modal data delivers rich and diverse information on the phenomenon relevant to the given task. Human is naturally born to be a good multi-modal learner in that it effectively learns from various modalities including audio, video, smell, touch, and so on. On the contrary, multi-modal fusion has been one of the most challenging problems in machine learning field due to the difficulty of combining the high level semantic information delivered by the different sources. Basically, multi-modal fusion concerns in which data processing stage the information fusion is conducted, which leads to the categorization into *early fusion* and *late fusion* [19]. While the early fusion aims to extract the joint representation directly from the raw or preprocessed data, the late fusion aggregates the decisions separately made by the machine learning models for each modality. The late fusion is considered to be easy to implement but its performance is limited in that the correlation structure underlying in multi-modal sources is not fully utilized. Early fusion is also difficult to find a good joint representation due to significantly different data structures between modalities. Recent emergence of deep neural network (DNN) technique (called deep learning) [18] has enabled the extraction of the hierarchical semantic features from the raw data and consequently led to better and flexible use of feature-level data fusion. The common practice for such feature-level fusion is to construct the shared representation by merging the intermediate features obtained by separate machine learning models. In this sense, this fusion approach is referred to as *intermediate fusion*. Leveraging the high level representation found by DNN, the *deep multi-modal learning* (DML) technique was shown to achieve remarkable performance for a variety of multi-modal learning problems including audio-visual speech recognition [21,23], multi-modal activity and emotion recognition [16,24,25], image analysis from RGBD data [7,10,11], and camera and Lidar sensor fusion [6,35].

The ultimate goal of the multi-modal learning is to achieve the highest level of reliability and robustness in performing the given task using the redundant information provided by multi-modal data. This implies that when the information provided by a single modality is not sufficiently good enough, the multi-modal learning uses the complementary information delivered by the different modalities and compensates for the performance degradation. The robustness against the degraded data quality can also be readily offered by the conventional late fusion approaches which aggregate the decisions in proportion to their credibility. On the contrary, it is not obvious how the intermediate fusion for DML can enjoy such selective information combining since it is difficult for the machine learning models to judge the reliability of the intermediate features. One conceivable approach is to train the fusion network with the data set containing various types of degradation, hoping that the architecture learns to use only reliable features from the multi-modal sources. However, our empirical evaluation reveals that the existing fusion architectures are not flexible enough to

adapt their fusion strategy to the variation in data quality. This quests the new DML architecture which can take the information as needed from each modality to achieve the robust performance.

This paper proposes the DML architecture that can offer robust performance for missing or degraded modalities. Towards this end, we introduce a feature-level gated information fusion (GIF) network which combines the features obtained for each modality in a way that only information relevant to the task is aggregated. The GIF network controls the amount of information flow incoming from each modality through *gating mechanism*. Specifically, the GIF network selectively gates the contribution of the features by weighting each element of features by the factor between 0 and 1. These weights are independently calculated through the separate network called weight generation (WG) network. The WG network takes all concatenated features for all modalities as an input and produces the weights by applying the convolution layers followed by the sigmoid function. In fact, this operation resembles the gating operations used in long short term memory (LSTM) [13] in that it controls the operation of information gating in a data-dependent manner. We build the deep 2D object detection architecture based on the proposed multi-modal fusion method. The proposed method first applies the multiple convolutional neural network (CNN) networks (e.g. VGG [31], ResNet [12], etc.) to generate the intermediate feature maps for the different modalities. Then, we combine these feature maps across the modalities through the proposed GIF network. The rest of the procedure to perform the object detection based on the joint feature maps found by the GIF network follows that of the single shot detector (SSD) [20].

The prior work most closely related to our work is [1], in which the similar gated fusion was used to extract the joint features from the text and image data. While the work in [1] focuses on the role of gating function for modality selection, we aim to highlight the different aspect of the gated fusion for improving the robustness of deep multi-modal fusion in the context of object detection. The key contributions of our work are highlighted below.

- We demonstrate that our gated fusion network can effectively improve the robustness of multi-modal learning. Note that developing a robust perception system using redundant sensors is a crucial problem in various safety-critical applications such as autonomous driving and mobile robot.
- We present the robust 2D object detector built upon the proposed multi-modal fusion scheme. The idea of our weighted information fusion is not limited to the object detection and can readily extended to other learning models utilizing multi-modal data.
- In order to promote the robustness of our scheme, we train our model using the special data augmentation strategy. We generate the additional examples by corrupting some of modalities in various ways (e.g. blanking, noise addition, occlusion, severe change in illumination) and guide our model to learn the way to fuse the different modalities with the proper weights.
- The experiments conducted with the SUN-RGBD dataset [32] and KITTI camera and Lidar dataset [8] show that the proposed architecture achieves

better detection accuracy than the baseline object detectors even when the subset of modalities are severely corrupted.

The rest of the paper is organized as follows. In Sect. 2, we review the previous literature on the DML. In Sect. 3, we present the details on the proposed GIF network and the robust 2D object detector based on multi-modal fusion. The experimental results are provided in Sect. 4 and the paper is concluded in Sect. 5.

2 Related Works

In this section, we briefly review the existing works on the DML methods.

2.1 Deep Multi-modal Learning

The earliest research on DML goes back to the works in [22] and [33] first showing that the effective joint data representation can be found using deep models such as deep autoencoder and deep Boltzman machine (DBM). Since then, the DML has been shown to work for a variety of learning tasks including representation learning, data fusion, translation, and alignment. (See [2] and [26] for comprehensive review on DML.) Among them, we specifically review the multi-modal data fusion due to the relevance to our work. The multi-modal fusion aims to extract as much relevant information on the task as possible from the data having heterogeneous characteristics. Since the emerging DNN is good at finding high-level semantic features through the hierarchical pipelined data processing, the intermediate fusion, which combines the features found at the middle layers of the DNN, has given rise to an effective means for multi-modal fusion.

Thus far, various DML techniques have been proposed for different types of modalities. In [22], the speech recognition was enhanced by using the joint data representation learned from the voice record and the video of lip movement. In [16], the audio feature from CNN and the visual features from the deep belief network were aggregated into single video descriptor for emotion recognition. In [21] and [24], the feature-level multi-modal fusion was shown to achieve good performance in the application to speech recognition and sentiment analysis, respectively. The DML architecture was also designed to generate the effective features for RGB-D (RGB-depth) and multi-view images. In [7], the feature vectors obtained from the fully connected (FC) layer of two separate CNNs were combined to generate the joint features for RGB-D images. In [10], the performance of the RGB-D fusion was improved by finding the effective encoding scheme for depth image. In [19], multi-level fusion architecture was proposed to learn multi-modal features for semantic segmentation.

2.2 Object Detection Using Multi-modal Data

Recently, the CNN led to remarkable performance improvement for the recognition of 2D image. Thus far, various CNN-based object detectors have been

proposed. Basically, these object detectors calculate the score for the bounding box candidate and the object class based on the feature map produced by the CNN. The state-of-the-art object detectors include the faster R-CNN [29], SSD [20], YOLO [27], and YOLOv2 [28]. The object detection can be extended for the multi-modal setup. In [36], object detection based on RGB-D data was performed using the cross-modality feature found by three CNN architectures. In [14], the deep fusion scheme based on RGB-D image was proposed using the *hallucination network* which learns a new RGB image representation by mimicking the depth network. In [6], the multi-view images are constructed from raw Lidar measurement data and used to perform 3D object detection along with RGB image in the context of automated driving. In [35], the authors proposed the *point-fusion network* which predicts the corner location of the 3D bounding box based on the Lidar 3D point data.

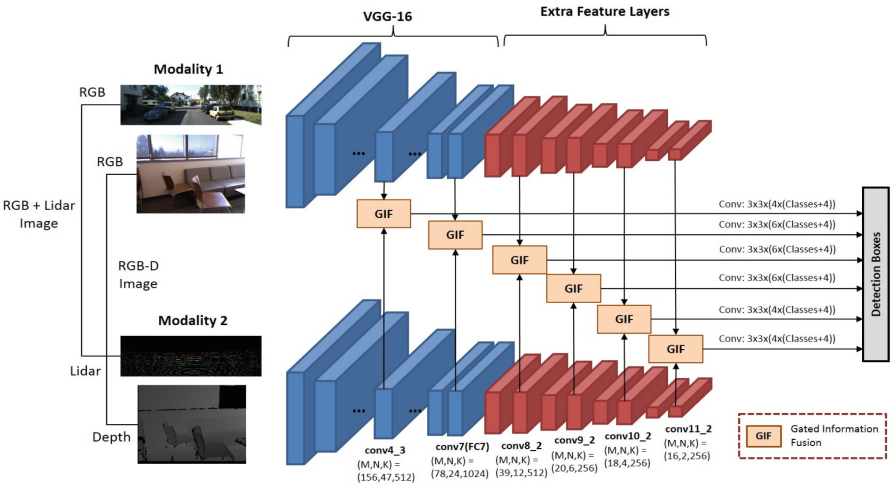


Fig. 1. Overall structure of the proposed R-DML. The R-DML takes the intermediate feature maps from both modality 1 and modality 2 using separate CNNs and combines them through the proposed GIF network. The joint feature maps produced by the GIF network are used to compute the score for object detection following the procedure of SSD.

3 Robust Deep Multi-modal Learning (R-DML)

In this section, we present our robust deep multi-modal learning (R-DML) architecture in details.

3.1 R-DML Architecture

Overall Architecture. The structure of the proposed R-DML is described in Fig. 1. Though our idea can be applied to the case of more than two modalities, we consider the example of two modalities. First, two separate CNN pipelines are used to extract the intermediate features to be fused. Each CNN consists of the CNN backbone network (e.g. VGG-16) followed by 8 extra convolutional layers. This configuration is similar to that of SSD. We combine the feature maps at the layers of conv4_3, conv7 (FC7), conv8_2, conv9_2, conv10_2, and conv11_2 layers.[1] These joint feature maps are used to perform object detection in different scales. As shown in Fig. 1, the GIF network is employed for feature-level information fusion. The GIF adjusts the contribution of the feature maps from each modality adaptively, whose detailed operation will be described next. In order to validate the benefit of the GIF, we compare our method with the baseline object detector referred to as the baseline DML (B-DML), which has the same structure as R-DML except that the combining weights in the GIF network are fixed to one.

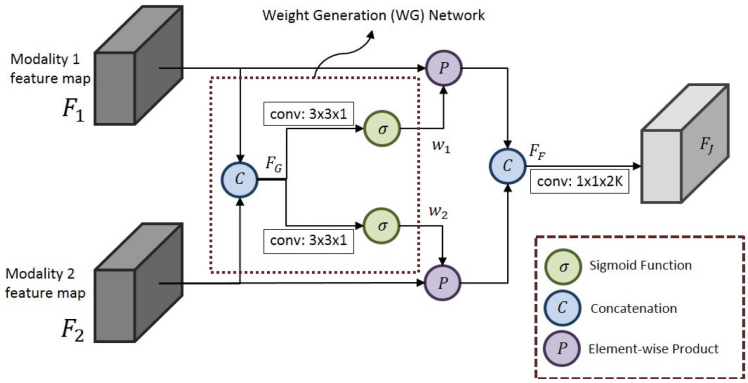


Fig. 2. The structure of the proposed GIF network. The GIF network produces the weight maps $\mathbf{w}_1$ and $\mathbf{w}_2$ by applying the convolutional layer and sigmoid function to the input features. Then, $\mathbf{w}_1$ and $\mathbf{w}_2$ are multiplied to the feature maps $\mathbf{F}_1$ and $\mathbf{F}_2$ for weighted information fusion.

Gated Information Fusion (GIF) Network. Figure 2 depicts the structure of the GIF network. The GIF network takes the intermediate feature maps from each CNN as an input and combines them with the weights calculated by the WG network. Let $\mathbf{F}_1$ and $\mathbf{F}_2$ be the $M \times N \times K$ feature maps obtained by two CNNs corresponding to two input modalities. The actual values of M, N and K for each layer are provided in Fig. 2. The GIF network consists of two parts: (1)

[1] We follow the notations of the SSD in [20].

the information fusion network and (2) the WG network. The information fusion network multiplies the $M \times N$ weight maps $\mathbf{w}_1$ and $\mathbf{w}_2$ to the feature maps $\mathbf{F}_1$ and $\mathbf{F}_2$, respectively. Such multiplication is done in element-wise for each feature map. Then, the weighted feature maps are concatenated across all modalities and $1 \times 1 \times 2K$ convolution is applied to fuse the feature maps. These operations result in the joint feature maps $\mathbf{F}_J$. Meanwhile, the WG network calculates the weights based on the input features as shown in Fig. 2. After concatenating the feature maps over all modalities, two separate $3 \times 3 \times 1$ CNN kernels $\mathbf{C}_1$ and $\mathbf{C}_2$ are applied in parallel to increase the depth in generate the high level features, which are used to calculate the combining weights[2]. Then, the sigmoid function is applied to produce the weight maps $\mathbf{w}_1$ and $\mathbf{w}_2$. We summarize the operation of the GIF network in the following equations.

$$\mathbf{F}_G = \mathbf{F}_1 \boxplus \mathbf{F}_2 \tag{1}$$

$$\mathbf{w}_1 = \sigma(\mathbf{C}_1 * \mathbf{F}_G + \mathbf{b}_1) \tag{2}$$

$$\mathbf{w}_2 = \sigma(\mathbf{C}_2 * \mathbf{F}_G + \mathbf{b}_2) \tag{3}$$

$$\mathbf{F}_F(i) = (\mathbf{F}_1(i) \odot \mathbf{w}_1) \boxplus (\mathbf{F}_2(i) \odot \mathbf{w}_2), \quad i = 1, ..., K, \tag{4}$$

$$\mathbf{F}_J = ReLU(\mathbf{C}_J * \mathbf{F}_F + \mathbf{b}_F) \tag{5}$$

where

- $\sigma(x) \triangleq \frac{1}{1+e^{-x}}$: sigmoid function(element-wise)
- $x * y$: convolutional layer
- $x \odot y$: element-wise product
- $x \boxplus y$: concatenation
- $\mathbf{F}(i)$: ith feature map of $\mathbf{F}$
- $\mathbf{b}_F, \mathbf{b}_1, \mathbf{b}_2$: biases of the convolution layers.

3.2 Training

Data Augmentation. In order to guide our network to learn to fuse the features appropriately in adverse environments, we design the data augmentation method. We generate the new training examples by applying various types of degradation to the subset of modalities. With such diverse training examples, our model would learn the robust multi-modal fusion. In our work, various type of modifications can be applied for data augmentation, including

- Blank Data (Type 1): we feed all pixel value to zero.
- Random occlusion (Type 2): we occlude the object using the black box whose size and location are randomly chosen.
- Severe illumination change (Type 3): we brighten the image in the rounded local region where the center and radius of the region and the brightness are randomly chosen.

[2] Our extensive experiments show that additional depth over single convolutional layer does not help improving the effectiveness of the gating operation.

– Additive random noise (Type 4): we add the random Gaussian noise where noise variance is randomly chosen within the certain range.
– No action.

The type of modification and which modalily will be degraded are chosen randomly with equal probability. Note that this data augmentation strategy is critical for our method to achieve the robust performance for the scenarios where some of modalities are corrupted.

Training Setup. Except for our data augmentation strategy, we use the same training setup used in SSD (e.g., matching strategy, hard negative mining, and multi-task loss function). We use VGG-16 pretrained model on ImageNet in two parallel CNN pipelines. The stochastic gradient descent (SGD) are used with the mini-batch size of 2 and the momentum parameter of 0.9. We set the initial learning rate to 0.0005. We set the weight decay parameter applied to L2 regularization term to 0.0005.

4 Experimental Results

In this section, we evaluate the performance of the proposed R-DML scheme using two public datasets: KITTI dataset [8] and SUN-RGBD dataset [32]. We first investigate the behavior of the gating operation to verify the effectiveness of the GIF network. Then, we compare the performance of our scheme with that of other multi-modal fusion schemes. Note that for fair comparison, we re-trained other algorithms using the same augmentation method as that used to train the R-DML. A total of 130 epochs and 200 epochs are executed with the KITTI dataset and SUN-RGBD dataset, respectively.

4.1 Datasets

KITTI Dataset. The KITTI dataset is collected by driving the car equipped with Pointgrey cameras and a Velodyne HDL-64E Lidar in various driving scenarios. The training set and test set contain 7,481 images and 7,518 images, respectively. Since the labels of the test images are not publicly available, we split the labeled training dataset into the training set and validation set by half as done in [5]. We evaluate the 2D detection performance with three object categories, i.e., car, pedestrian, and cyclist and three difficult levels, i.e., easy, moderate, hard as proposed in the KITTI Benchmark.

We consider the multi-modal fusion task which performs object detection using both RGB image and 3D lidar data. In order to preprocess the data for our object detector, we convert the 3D point cloud data into the 2D image in camera plane. The 3D point data in KITTI dataset contains the 3D coordinate (X, Y, Z) and the reflectivity R measured for each reflected laser pulse. Specifically, we

map the 3D coordinate (X, Y, Z) of Lidar data into the 2D coordinate (x, y) on camera plane using

$$\begin{bmatrix} x \\ y \end{bmatrix} = calib\_matrix \cdot \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}. \tag{6}$$

where *calib_matrix* is the matrix for coordinate transformation. Note that we quantize (x, y) to the nearest integer and limit the maximum range of (x, y) by that of camera plane. For the given 2D coordinate (x, y), we create three channel image by encoding the values of X, Z, and R to the pixel values. This creates the image with the depth, height, and intensity (DHI) channels. The pixel values for the DHI channels are obtained by

$$val_d = 255 \cdot (1 - \min[X/max\_X, 1]) \tag{7}$$

$$val_h = 255 \cdot (1 - \min[Z/max\_Z, 1]) \tag{8}$$

$$val_i = 255 \cdot (1 - \min[R/max\_R, 1]). \tag{9}$$

Note that $X \in [0, max\_X]$, $Z \in [0, max\_Z]$, and $R \in [0, max\_R]$ are mapped to the pixel values between $[0, 255]$ in a linear scale. For example, we set max_X, max_Z, and max_R to 80 m, 6 m, and 0.7. Note that the DHI Lidar image and the RGB camera image of the size 1242×375 are used as the multi-modal inputs for the proposed object detector. We apply data augmentation to these images. Since it is hard to introduce noise and illumination change to the Lidar image, we apply them only for RGB image.

(a) RGB image (b) Illumination (c) Noise (d) Occlusion

Fig. 3. Examples of modifications applied to the camera image on SUN-RGBD dataset.

SUN-RGBD Dataset. The SUN-RGBD dataset is a large-scale RGB-D dataset collected in indoor environments. It contains 10,335 RGB image and depth image pairs including NYUDv2 depth [30], Berkeley B3DO [15], and SUN3D [34]. The dataset consists of 5,285 training set and 5,250 testing set. We evaluate the detection performance with 19 object categories as in [32]. Note that we apply the same data augmentation strategy used for the KITTI dataset and we set the size of the input image to 530×400. The examples of the modifications applied to the RGB camera image in SUN-RGBD dataset are illustrated in Fig. 3.

Extended Test Dataset. To evaluate the robustness of the proposed R-DML, we randomly generate the test dataset using the same types of data modification applied for the data augmentation. Both KITTI and SUN-RGBD datasets contain the RGB camera image while the modality 2 corresponds to the Lidar image and depth image, respectively. In our experiments, we come up with the following test cases:

- Total: Test with all normal and degraded examples together.
- RGB+modality2: Test with the normal test examples without any degradation.
- RGB (blank)+modality2: Test with the test examples with RGB image blanked.
- RGB+modality2 (blank): Test with the test examples with modality 2 blanked.
- RGB (occlusion)+modality2: Test with the test examples with RGB image occluded.
- RGB+modality2 (occlusion): Test with the test examples with modality 2 occluded.
- RGB (noise)+modality2: Test with the test examples with the noise of the RGB image changed.
- RGB (illumination)+modality2: Test with the test examples with the illumination of the RGB image changed.

Note that the performance evaluation is performed with the same number of test examples for each case.

4.2 Experimental Results on KITTI Dataset

First, we evaluate the performance of the proposed method when tested on KITTI dataset. As a baseline algorithm, the following multi-modal fusion methods are considered:

- B-DML: It has the same structure with R-DML except that both gating weights applied to two modalities are fixed to one.
- Early fusion: We concatenate two modality inputs and feed them into a single SSD.
- SSD-based fusion: We take the late fusion approach, which combines the detection boxes generated by two SSDs. Both SSDs are trained with the camera and Lidar images, separately. We combine the detection boxes found by two SSDs using non-maximum suppression.
- AVOD [17]: This is the state-of-the-art multi-modal object detection algorithm using both camera and Lidar data. Though the AVOD is capable of 3D object detection, we transform the 3D box information into the front view to compare it with our method.

Table 1 provides the average precision (AP) achieved by the algorithms of interest for *Car* category. The proposed data augmentation strategy is used for

Table 1. Detection performance (AP) for car category on the extended KITTI test dataset

Test input	Our R-DML			B-DML			Early fusion			SSD-based fusion [20]			AVOD [17]		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
Total	**93.95**	**86.70**	**78.05**	89.86	82.21	72.21	91.10	85.65	75.83	89.69	82.03	72.96	-	-	-
RGB + Lidar	**98.69**	**90.31**	**82.16**	93.61	87.01	77.52	95.84	89.94	79.67	91.72	87.93	78.46	89.85	87.99	80.27
RGB (blank) + Lidar	88.86	78.12	69.68	86.56	74.30	64.71	**89.94**	**78.99**	69.56	87.92	77.83	69.11	86.42	69.82	**69.77**
RGB + Lidar (blank)	**97.39**	**90.29**	**81.84**	91.88	88.10	78.68	90.48	88.56	77.92	93.31	89.27	80.03	-	-	-
RGB (occl.) + Lidar	89.88	88.12	**79.03**	88.12	78.52	68.85	90.22	84.15	73.93	**91.78**	**88.22**	78.80	87.94	78.75	78.53
RGB + Lidar (occl.)	**97.72**	**90.23**	**81.94**	92.75	87.10	77.67	90.53	88.91	79.07	84.80	74.88	66.33	-	-	-
RGB (noise) Lidar	89.33	80.15	71.12	86.75	75.13	65.71	**90.18**	**81.29**	72.04	88.67	76.12	67.18	88.88	79.79	**79.46**
RGB (illum.) + Lidar	**95.82**	**89.71**	**80.58**	89.37	85.31	75.87	90.48	88.42	78.60	89.69	79.96	70.82	88.60	79.33	79.00

Table 2. Detection performance (AP) for car category on the extended KITTI test dataset with unseen types of modification

Test input	Our R-DML			B-DML		
	Easy	Mod.	Hard	Easy	Mod.	Hard
RGB + Lidar (Type1. blank)	-	-	-	-	-	-
RGB + Lidar (Type2. occl.)	**83.31**	**82.23**	**74.41**	80.50	77.37	68.89
RGB + Lidar (Type3. illum.)	**90.62**	**89.06**	**80.04**	89.70	87.22	78.59
RGB + Lidar (Type4. noise)	**83.10**	**73.34**	**65.67**	78.15	66.52	58.25

training all methods considered. The AP is evaluated using 3,740 test examples for each scenario. We observe that the proposed R-DML shows better detection accuracy than other algorithms in almost all cases. In particular, the R-DML significantly outperforms the B-DML, which shows the benefit of the proposed gated fusion method. We see that the performance gain of the R-DML over B-DML can go up to 5% of AP for some test scenarios (e.g. occlusion case). Interestingly, the proposed scheme outperforms the B-DML even when the normal KITTI data is used without any data corruption for test. Since this KITTI dataset might contain some natural but somewhat benign level of real world perturbation (e.g. camera noise and adverse illumination change), this could be a part of evidence showing that the R-DML is robust to real world perturbation as well as synthetic one. In essence, all these results show that the proposed GIF

Table 3. Detection performance (AP) on KITTI validation set. (*: trained by us, red text: ranked first, blue text: ranked second, green text: ranked third)

Method	Data	Easy	Moderate	Hard
SSD* [20]	Mono	93.31	89.27	80.03
3DOP [5]	Stereo	94.49	89.65	80.97
Mono3D [4]	Mono	95.75	90.01	80.66
Deep Manta [3]	Mono	97.58	90.89	82.72
MV3D [6]	Lidar+Mono	95.01	87.59	79.90
SSD-based fusion*	Lidar+Mono	91.72	87.93	78.46
B-DML*	Lidar+Mono	93.61	87.01	77.52
Our R-DML*	Lidar+Mono	98.69	90.31	82.16

network provides better model flexibility to improve the performance of multi-modal fusion. We evaluate the detection performance on *Pedestrian* and *Cyclist* categories as well. We obtain 70.59 (R-DML) versus 68.37 (B-DML) for moderate level for the pedestrian category and 70.11 (R-DML) versus 68.90 (B-DML) for the cyclist category. The whole results are provided in the supplemental material.

In Table 1, we have tested the models using the same type of modification used for training. However, the real world perturbation is hard to predict so that it is impossible to synthesize it in the training phase. Thus, the additional experiments are designed to evaluate how well the proposed method generalizes to the unseen types of modification. We train the models using the data augmented with the type 1 to $(i-1)$ modification and then test with the type i modification. For example, the models trained with the dataset augmented by the type 1 (blanking) and type 2 (occlusion) modification are tested with the type 3 (illumination change) modification. Table 2 shows that the R-DML achieves the performance gain over the B-DML when tested with each degradation type. This shows that the proposed method exhibits the robust behavior when encountered with unseen types of degradation.

Next, we look into the behavior of the gating operation in details. Figure 4 shows the histogram of the GIF weights (averaged over the whole weight map at the conv4_3 layer) for the case that the RGB image is completely blanked. Note that the weights multiplied to the RGB features are close to zero in order to reduce the contribution from the blanked data. On the contrary, we see that the weights for the normal Lidar image are close to one. In Fig. 5, we visualize the GIF weight maps learned by the GIF for the case where the RGB image is locally occluded by the black box. We find that the GIF weights in the camera side are small only within the locally occluded region while they are high for the rest of area. Note that the GIF weights for the Lidar image are relatively high for the whole region. This shows our gating mechanism controls the amount of information combined depending on the quality of the features for each interested region.

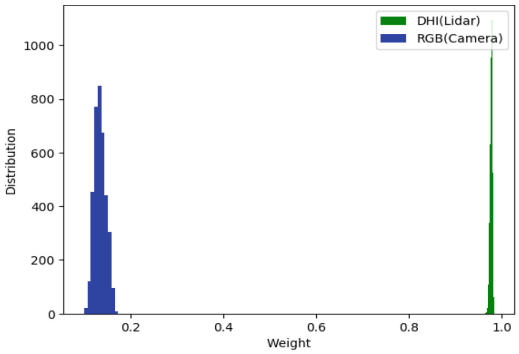
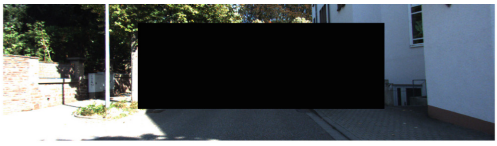
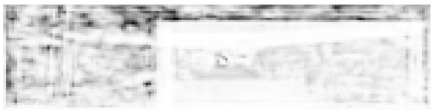


Fig. 4. The histogram of the averaged GIF weights at conv4_3 layer. The weights for the blanked data are close to zero. This demonstrates the operation for reducing the contribution from unreliable data.



(a) The locally occluded RGB image for test

(b) The weight map applied to RGB feature maps at conv4_3 layer



(c) The weight map applied to the Lidar feature maps at conv4_3 layer

Fig. 5. The visualization of the GIF weight maps at conv4_3 layer. Note that the weights for the RGB features are reduced significantly for the occluded region. This shows that the gating operation conducts locally controlled information fusion.

In Table 3, we compare the performance of the proposed method with other state of the art 2D object detectors when tested with the original KITTI dataset. The candidate detectors include SSD [20], 3DOP [5], Mono3D [4], Deep Manta [3], and MV3D [6]. For fair comparison, we use the same train/validation split method used in [3–6]. Note that SSD-based fusion, B-DML and the proposed R-DML are trained with the proposed data augmentation schemes. We observe that the performance of the proposed object detector is better or on par with the other algorithms for all difficulty levels. This shows that the proposed fusion method exhibits competitive performance for the normal environment while promising the robust performance in the adverse environment. Note that even though the proposed R-DML is built upon the baseline SSD, significant performance gain is achieved over the baseline SSD through the multi-modal fusion strategy proposed in our work. It is interesting that the B-DML does not perform better than the

SSD. This issue appears to be due to different data augmentation strategies used for training the B-DML and SSD. Due to the limitation of the SSD taking only single input modality, we could not train the SSD with our data augmentation strategy. On the other hand, the B-DML is trained with the data augmentation. We see that the B-DML does not achieve better performance than the SSD with the normal KITTI data. On the contrary when both methods are trained without data augmentation, the B-DML outperforms the SSD. Note that our R-DML significantly outperforms the B-DML and the SSD for both normal and extended KITTI datasets.

4.3 Experimental Results on SUN-RGBD Dataset

Table 4 provides the mean average precision (mAP) of the proposed object detection algorithm. Since there are not many recent 2D object detection algorithms using SUN-RGBD dataset, we compare our method with only the B-DML and supervision transfer (ST) methods [11]. The ST method is the fast-RCNN [9] based object detector which combines the detection boxes obtained by two separate object detectors. For fair comparison, we trained the B-DML and ST with the same augmentation method as that used for our R-DML. For each test case, we use the 5,250 test examples. We see in Table 4 that the proposed R-DML achieves better detection accuracy than the B-DML, which reveals the effectiveness of our gated fusion algorithm for the task of the RGB and depth image fusion as well. Note that the R-DML maintains the performance gain over the B-DML even when the normal SUN-RGBD dataset are used for test without any modification. The AP results per category are provided in the supplemental material.

Table 4. Results for detection performance (mAP) on extended SUN-RGBD test dataset

Test Input	Our R-DML	B-DML	Supervision transfer [11]
Total	**34.72**	29.13	21.35
RGB + depth	**40.43**	36.31	26.68
RGB (blank) + depth	**30.76**	28.37	15.81
RGB + depth (blank)	**32.69**	12.03	22.25
RGB (occlusion) + depth	**35.65**	29.39	22.95
RGB + depth (occlusion)	**35.04**	33.12	22.75
RGB (noise) + depth	**32.76**	31.50	16.65
RGB (illumination) + depth	**35.67**	33.19	22.40

5 Conclusions

In this paper, we proposed the robust multi-modal learning technique which fuses the intermediate features produced by the CNN with appropriate contributions. Inspired by the gating mechanism used in LSTM, we devised the gated information fusion network, which combines the features from each modality with the weights computed based on the input features to be fused. Such GIF network was used to perform 2D object detection using multi-modal inputs and the whole system is trained end-to-end. We used the special data augmentation strategy for promoting the robustness of our system, which corrupts some of modalities using various artificial operations. The experiments performed over KITTI dataset and SUR-RGBD dataset shows the superiority of the proposed method for the scenarios of missing or degraded modalities.

References

1. Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992 (2017)
2. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 423–443 (2018)
3. Chabot, F., Chaouch, M., Rabarisoa, J., Teulière, C., Chateau, T.: Deep manta: a coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image. In: Proceedings of IEEE Conference on Computer Vision Pattern Recog (CVPR) (2017)
4. Chen, X., Kundu, K., Zhang, Z., Ma, H., Fidler, S., Urtasun, R.: Monocular 3D object detection for autonomous driving. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
5. Chen, X., et al.: 3D object proposals for accurate object class detection. In: Advance in Neural Information Processing Systems (2015)
6. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3D object detection network for autonomous driving. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
7. Eitel, A., Springenberg, J.T., Spinello, L., Riedmiller, M.A., Burgard, W.: Multimodal deep learning for robust RGB-D object recognition. In: Proceedings of IEEE/RSJ Interernational Conference on Intelligent Robots and Systems (IROS) (2015)
8. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The kitti vision benchmark suite. In: Proceedings of IEEE Confernce on Computer Vision and Pattern Recognition (CVPR) (2012)
9. Girshick, R.: Fast R-CNN. In: Proceedings IEEE International Conference on Computer Vision (ICCV) (2015)
10. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from RGB-D images for object detection and segmentation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 345–360. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10584-0_23
11. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
14. Hoffman, J., Gupta, S., Darrell, T.: Learning with side information through modality hallucination. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
15. Janoch, A., et al.: A category-level 3D object dataset: putting the kinect to work. In: Fossati, A., Gall, J., Grabner, H., Konolige, K., Ren, X. (eds.) Consumer Depth Cameras for Computer Vision. ACVPR, pp. 141–165. Springer, London (2013). https://doi.org/10.1007/978-1-4471-4640-7_8
16. Kahou, S.E., et al.: Emonets: multimodal deep learning approaches for emotion recognition in video. J. Multimodal User Interfaces **10**, 99–111 (2015)
17. Ku, J., Mozifian, M., Lee, J., Harakeh, A., Waslander, S.: Joint 3D proposal generation and object detection from view aggregation. arXiv preprint arXiv:1712.02294 (2017)
18. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015)
19. Li, Y., Zhang, J., Cheng, Y., Huang, K., Tan, T.: Semantics-guided multi-level RGB-D feature fusion for indoor semantic segmentation. In: 2017 IEEE International Conference on Image Processing (ICIP) (2017)
20. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
21. Mroueh, Y., Marcheret, E., Goel, V.: Deep multimodal learning for audio-visual speech recognition. In: Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) (2015)
22. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of International Conference on Machine Learning (ICML) (2011)
23. Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H.G., Ogata, T.: Audio-visual speech recognition using deep learning. Appl. Intell. **42**(4), 722–737 (2015)
24. Poria, S., Cambria, E., Gelbukh, A.: Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proceedings of Conference Empirical Methods in Natural Language Processing, pp. 2539–3544 (2015)
25. Radu, V., Lane, N.D., Bhattacharya, S., Mascolo, C., Marina, M.K., Kawsar, F.: Towards multimodal deep learning for activity recognition on mobile devices. In: Proceedings of 2016 ACM Interernational Joint Confernce on Pervasive and Ubiquitous Computing, pp. 185–188 (2016)
26. Ramachandram, D., Taylor, G.W.: Deep multimodal learning. IEEE Signal Process. Mag. **34**(6), 96–108 (2017)
27. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of IEEE Confernce on Computer Vision and Pattern Recognition (CVPR) (2016)
28. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (2015)

30. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 746–760. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_54
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
32. Song, S., Lichtenberg, S.P., Xiao, J.: Sun RGB-D: a RGB-D scene understanding benchmark suite. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
33. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep boltzmann machines. J. Mach. Learn. Res. **15**, 2949–2980 (2014)
34. Xiao, J., Owens, A., Torralba, A.: Sun3D: a database of big spaces reconstructed using SFM and object labels. In: Proceedings of IEEE International Conference on Computer Vision (ICCV) (2013)
35. Xu, D., Anguelov, D., Jain, A.: Pointfusion: deep sensor fusion for 3D bounding box estimation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
36. Xu, X., Li, Y., Wu, G., Luo, J.: Multi-modal deep feature learning for RGB-D object detection. Pattern Recogn. **72**, 300–313 (2017)