
MMRNet: Improving Reliability for Multimodal Object Detection and Segmentation for Bin Picking via Multimodal Redundancy

Yuhao Chen¹, Hayden Gunraj¹, E. Zhixuan Zeng¹,
Robbie Meyer¹, Maximilian Gilles², Alexander Wong¹

¹ University of Waterloo, Canada

² Karlsruhe Institute of Technology, Germany

{yuhao.chen1, hayden.gunraj, ezzeng, robbie.meyer, alexander.wong}@uwaterloo.ca,
maximilian.gilles@kit.edu

Abstract

Recently, there has been tremendous interest in industry 4.0 infrastructure to address labor shortages in global supply chains. Deploying artificial intelligence-enabled robotic bin picking systems in real world has become particularly important for reducing stress and physical demands of workers while increasing speed and efficiency of warehouses. To this end, artificial intelligence-enabled robotic bin picking systems may be used to automate order picking, but with the risk of causing expensive damage during an abnormal event such as sensor failure. As such, reliability becomes a critical factor for translating artificial intelligence research to real world applications and products. In this paper, we propose a reliable object detection and segmentation system with **MultiModal Redundancy** (MMRNet) for tackling object detection and segmentation for robotic bin picking using data from different modalities. This is the first system that introduces the concept of multimodal redundancy to address sensor failure issues during deployment. In particular, we realize the multimodal redundancy framework with a gate fusion module and dynamic ensemble learning. Finally, we present a new label-free multimodal consistency (MC) score that utilizes the output from all modalities to measure the overall system output reliability and uncertainty. Through experiments, we demonstrate that in an event of missing modality, our system provides a much more reliable performance compared to baseline models. We also demonstrate that our MC score is a more reliability indicator for outputs during inference time compared to the model generated confidence scores that are often over-confident.

1 Introduction

Global labor shortages and the need for resilient supply chains has accelerated companies' upgrades to industry 4.0 and introduced a range of technologies such as big data, cloud computing, internet of things (IoT), robotics, and artificial intelligence (AI) into production systems. With warehouses and manufacturing units becoming smart environments, a crucial objective is to develop an autonomous flow of both material and information, and robotic bin picking plays an essential role in this task.

Robotic bin picking has been an active area of research for many decades given the complexity of the task, ranging from joint control and trajectory planning [15] to object identification [1] and grasp detection [4]. In particular, we examine the object detection and segmentation task in autonomous bin picking. Different from object detection and segmentation in other areas such as autonomous driving, robotic vision system works in environments that are very close to the camera, dealing with heavy

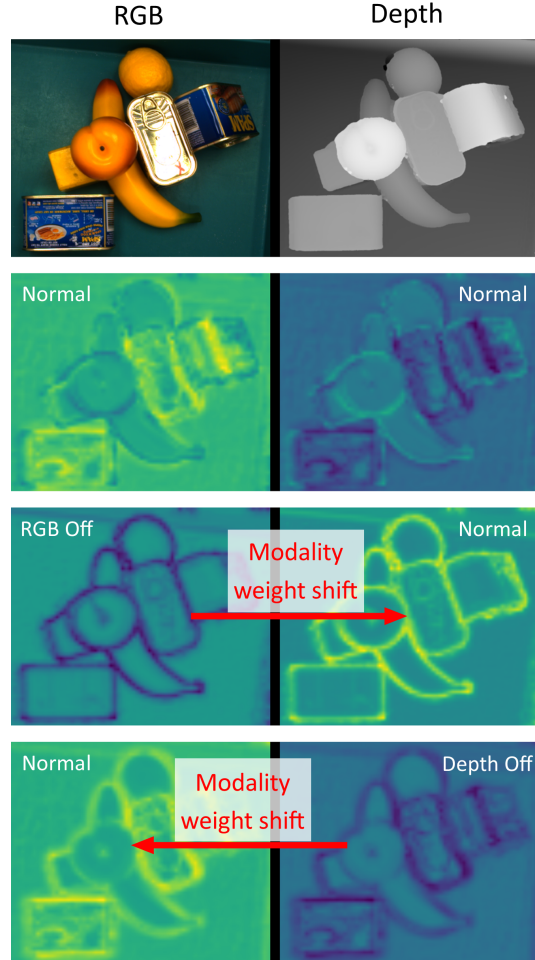


Figure 1: The dynamic modality weight shifting of our network ensures a reliable overall performance when a modality is missing. Row 2-4 heatmaps describe the average gate weights of each modality at a single feature scale. Yellow indicates high weight, dark purple indicates low weight.

occlusions, shadows, dense object layouts, and complex stacking relations. It plays an essential role in a robot’s perception system.

Deep neural networks have been proven effective for object detection and segmentation [13, 24]. But, deploying such systems in robotic picking applications is challenging due to the many sources of uncertainty present in practical scenarios. Real-world bin scenes may consist of a wide variety of unknown or occluded items arranged in an infinite number of poses and illuminated with variable lighting conditions. In addition to the variability of real-world bin scenes, errors in the camera system can make a computer vision system unreliable. Camera sensors are prone to noise and can fail in various situations such as specular reflections (missing values), black areas (missing depth), overexposure, blur, and artifacts. In practice, commercial systems are expected to run 24/7 to be feasible, which increases the risk of imaging sensor failures compared to research environments. If not accounted for, sensor failures can lead to wrongly commissioned orders and in the worst case to product and hardware damages, leading to expensive recall campaigns or production downtime. Therefore, vision systems capable of handling uncertain inputs and producing reliable predictions under sensor errors are critical to creating fail-safe applications.

One approach to creating fault-tolerant object detection and segmentation systems is to introduce system duplication, where portions of the system are duplicated to allow the system to continue to operate despite failures of its constituent parts. This approach assumes that failures are caused by either input sensor failures or computational failures. However, duplication may not provide

fault-tolerance in situations where the system is operating correctly but its sensors are unable to adequately measure the inputs. For example, a camera may fail to adequately image a piece of glass due to its transparency, and so the use of a second identical camera cannot address this issue. In addition, deep neural networks as a data-driven approach are designed to capture feature distributions of the input dataset. A simple duplication of these networks will not detect features that are not in the training distribution. Instead, we add image data from depth sensor as an additional modality to capture object feature characteristics from a different perspective. More specifically, depth data has very simple texture yet rich geometric features, that are more transferable to unseen objects than RGB data.

A good system duplication design duplicates components that are more likely to fail, preventing any disruption in the information flow from the system input to output. Non-data-driven methods have well defined explicit logic to control the information flow. In comparison, deep learning system learns the input and output mapping through high-dimensional implicit feature representations. A typical deep learning model encodes input information through a backbone network into a high-dimensional latent representation, and downstream tasks use the representation to predict low-dimensional outputs. Consequently, a large amount of information is lost during the dimensionality reduction of downstream tasks. However, in robotic bin-picking, unseen items may contain highly complex image characteristics that require both RGB and depth to work collaboratively. For example, RGB backbones are better at detecting transparent objects and depth backbones are better at detecting dark objects. A pair of eyeglasses with black frame will require the RGB backbone to focus on glass parts while the depth backbone to focus on the frame for a complete detection and segmentation. With the reduction of dimensionality, a simple result aggregation on two low quality detections will create another low quality detection. Additional result merging networks or explicit merging logic will introduce errors and instabilities into the system. An effective modality fusion technique that will dynamically fuse modality features with limited loss of information is therefore greatly desired. In addition, modality features merging may introduce dependencies between them, causing unexpected model behavior when one of the modality feature is absent. We tackle this problem with a multimodal redundancy framework consists of two key techniques: 1) we use a multi-scale soft-gating mechanism to make the network learn to weigh and combine features between modalities dynamically, and 2) we use a dynamic ensemble learning strategy to train the sub-system independently and collaboratively in an alternating fashion. With this framework, only one modality needs to be present for the model to operate.

Finally, we propose a novel multimodal consistency (MC) score as a more objective reliability indicator for the system output based on the overlaps of detected bounding boxes and segmentation masks. This can be used as an indicator for model uncertainty on individual predictions, as well as model reliability on particular datasets.

Through experiments, we demonstrate that in an event of missing modality, our MMRNet provides a much more reliable performance compared to baseline models. When depth is removed, our network’s performance drop is within 1% where other models have a performance drop greater than 6%. When RGB is removed, our network’s performance drop is within 11% where other models have a performance drop greater than 80%. Furthermore, we demonstrate that our MC score is a more reliable indicator for output confidence during inference compared to the often overly-confident confidence scores. We summarize our contribution as the following:

- A multimodal redundancy framework consisting of a multi-scale soft-gating feature fusion module and a dynamic ensemble learning strategy allowing trained sub-systems to operate both independently and collaboratively.
- A multimodal consistency score to describe the reliability of the system output.

2 Related Work

Reliability study for deep learning-based systems: Deep learning-based methods are data-driven, encoding the decision making process through continuous latent vectors, which makes the model behavior hard to predict and fix. Only a few of the studies focus on the reliability aspect of the deep learning-based systems. In [26], Santhanam *et al.* list differences between traditional and deep learning-based software systems and discuss the challenges involved in the development of reliable deep learning-based systems. In [31], Xu *et al.* study the reliability of object detection systems in

autonomous driving. In [7], dos Santos *et al.* study the relationship between reliability and GPU precision (half, single, and double) for object detection tasks. Other reliability related work can be found in model uncertainty estimation [10]. To the best of our knowledge, none of the work investigates reliability or uncertainty for multimodal applications, in particular for robotic bin picking.

Multimodal Data Fusion: Multimodal learning [22, 17, 23, 8, 14] has been rigorously studied. In multimodal learning, there are three types of data fusion: early fusion, intermediate fusion, and late fusion. Each corresponds to merging information at input, intermediate, and output stage respectively. Early fusion involves combining and pre-processing inputs. A simple example is replacing the blue channel of RGB with depth channel [18]. Late fusion merges the low-dimensional output of all networks. For example, Simonyan *et al.* [27] combine spatial and temporal network output with i) averaging, and ii) linear Support Vector Machine [28]. Early fusion and late fusion are simpler to implement but have a lower dimensional representation compared to the intermediate fusion. Intermediate fusion involves merging high-dimensional feature vectors. Common intermediate fusion includes concatenation [22], and weighted summation [1]. Recently, more advanced techniques are developed to dynamically merge the modalities. In [29], Wang *et al.* propose a feature channel exchange technique based on Batch Normalization’s [16] scaling factor to dynamically fuse the modalities. In [5], Cao *et al.* propose to replace the basic convolution operator with Shapeconv to achieve RGB and depth fusion at the basic operator level. In [32], Xue *et al.* focus on the efficiency aspect of multimodal learning and propose a hard gating function which outputs an one-hot encoded vector to select modalities. In robotic grasping, Back *et al.* [1] take the weighted summation approach and propose a multi-scale feature fusion module by applying a 1x1 convolutional layer to the feature layers before passing them into a feature pyramid network (FPN) [19].

The aforementioned works are designed to optimize the overall network performance but at the same time introduce dependencies among modality features, which are extremely vulnerable in case of an abnormal event, such as an input sensor failure. In this paper, we address the multimodal fusion strategy from the system reliability perspective, where our goal is to design a simple yet effective network architecture that enables sub-modal systems to work independently as well as collaboratively to increase the overall system reliability.

Ensemble learning: Ensemble learning typically involves training multiple weak learners and aggregating their predictions to improve predictive performance [35]. One of the simplest approaches to construct ensembles is bagging [3], where weak learners are trained on randomly-sampled subsets of a dataset and subsequently have their predictions combined via averaging or voting techniques [35]. Instead of aggregating predictions directly, one may also use a meta-learner which considers the input data as well as each weak learner’s predictions in order to make a final prediction, a technique known as stacking [30]. Boosting [9] is another common approach where weak learners are added sequentially and leverage the previous learner’s mistakes to re-weight training samples, effectively attempting to correct the previous learner’s mistakes.

While ensemble learning has long been a common technique in classical machine learning, it can be expensive to apply to deep learning due to the increased computational complexity and training time of deep neural networks. Of particular relevance to this work is the application of ensemble learning to multimodal deep learning problems. In multimodal problems, the data distributions typically differ significantly between modalities and thus may violate the assumptions of certain ensembling techniques [20]. Nevertheless, ensemble methods have been applied to a variety of multimodal problems [20, 21, 6, 34]. For example, Menon *et al.* [21] trained modality-specific convolutional neural networks on three different magnetic resonance imaging modalities and combined the models’ predictions via majority voting. In [34], Zhou *et al.* used a stacking-based approach to combine the outputs of neural networks trained on text, audio, and video inputs, thereby reducing noise and inter-modality conflicts.

Rather than combining multiple models with a typical ensembling strategy, in this work we consider a *dynamic* ensemble where multiple unimodal systems are dynamically fused into a single network. This network is capable of both unimodal operation using each of its inputs independently as well as multimodal operation through the fusion of the constituent unimodal systems.

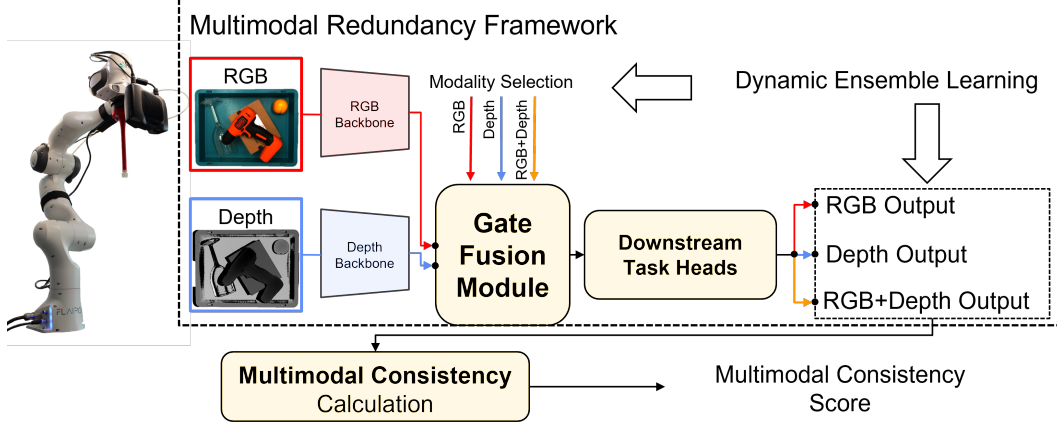


Figure 2: Block diagram of our multimodal redundancy framework. Gate fusion module allows simple switching between modalities. Trained with dynamic ensemble learning, our system is able to use both modalities independently (RGB or depth output) as well as collaboratively (RGB+depth output). A multimodal consistency score is computed at the end to indicate the reliability of the output.

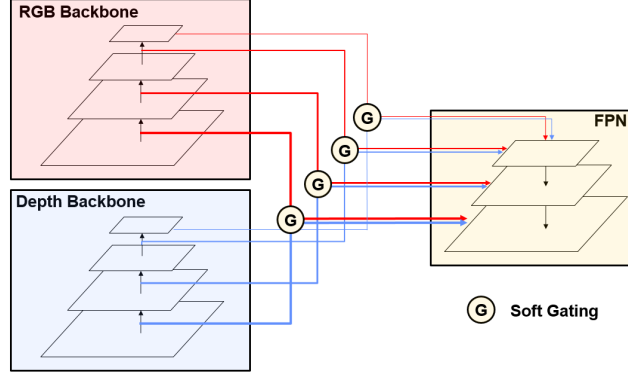


Figure 3: Gate fusion module fuses the multi-scale feature from each modality.

3 Methodology

The subsequent sub-sections outline the key components of our MMRNet architecture. Firstly, we introduce a multi-scale soft gating mechanism that effectively combines information from the two modalities. Secondly, we propose a dynamic ensemble learning strategy, which, in conjunction with the multi-scale soft gating mechanism, constitutes the multimodal redundancy framework. This framework helps to remove the inter-modality dependencies. Lastly, we present the formulation of the multimodal consistency score, which serves as our system’s reliability measure. We show our system block diagram in Figure 2.

3.1 Multimodal Redundancy Framework

Multi-scale Soft-Gate Feature Fusion (MSG Fusion): Fusing high-dimensional latent representation from two data distribution involves integrating information from multiple scales as well as multiple modalities. While a simple convolution as proposed in [1] can merge the information, it also constrains the information exchange between modalities to be within the same scale. The other modality’s high-level features may contain crucial contextual information for localizing and segmenting objects with intricate RGB and depth features. In order to maximize the utilization of contextual information from both modalities, we concatenate the features and input them into a Feature Pyramid Network (FPN) [19]. This FPN fuses multi-scale modality features in a hierarchical manner, enabling effective contextualization. Nonetheless, this process can result in inter-modality dependencies. To address this issue, we draw inspiration from [33] and incorporate a soft gating

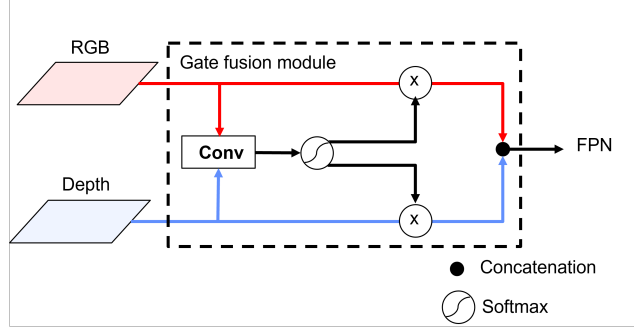


Figure 4: Soft gating architecture applied to every scale of feature layers.

mechanism. This mechanism enables the dynamic adjustment of feature weights from each backbone, thereby facilitating modality feature selection that is optimized for detecting individual object classes. More importantly, this method enables the model to disentangle features from different modality backbones. We define the total number of modalities to be N and denote the j th scale feature layer in modality m as $f_{m,j}$. Features of all modalities pass through a 1×1 convolution layer G_m . The convolution layer takes N j th scale modality features with C channels and outputs one feature layer with C channels for modality m . We obtain $g_{m,j}$:

$$g_{m,j} = G_m(\{f_{m,j} | m \in [0, N]\}) \quad (1)$$

The output gate weight $w_{m,j}$ is calculated by:

$$w_{m,j} = \sigma(\{g_{m,j} | m \in [0, N]\}) \quad (2)$$

, where σ is the softmax function ensuring modality weights sum to one. Finally, the gated feature layer for scale j and modality m is updated by:

$$f_{m,j} \leftarrow f_{m,j} w_{m,j} \quad (3)$$

We show the gate fusion module architecture in Figure 3 and Figure 4.

Dynamic Ensemble Learning Strategy: Although the proposed soft gating mechanism enables dynamic re-weighting of features extracted from each input modality, it does not inherently allow for modalities to be used independently. Ideally, the network would be capable of operating reasonably using a single input modality, with each additional modality providing improved performance or reliability. This accounts for the practical scenario where the sensor used to capture an input modality fails, forcing the system to leverage its other inputs.

Classic ensemble approaches combine weak models according to their standalone performance by a simple discrete process such as weighted sum [35]. In comparison, our gating module allows more dynamic interaction since information can be exchanged across different modalities and scales with respect to different input items, poses, and scene layouts. Instead of independently training each modality model and combine them with a weighted sum, we propose a novel dynamic ensemble learning strategy to train multimodal deep learning models, allowing for different modalities to be used both collaboratively and independently.

Specifically, in each training iteration we randomly select one of the possible input conditions: both inputs, RGB-only input, or depth-only input. In the unimodal conditions, we force the system to make predictions with only one of its usual inputs in order to encourage rich features to be extracted from both modalities. This training scheme prevents the model from learning to rely heavily on a single modality while simultaneously allowing the model to learn how to combine data from both modalities.

3.2 Multimodal Consistency Score

Existing object detection and segmentation networks contain a confidence score calculated by the softmax classifier for each detection. The reliability information in this score is somewhat subjective as it is estimated from the same network. Instead, we leverage the multimodal property of our model. In an ideal scenario, if we train a separate model for each modality, all models would converge to produce the same output describing the same object in the physical space. Less reliable models

will produce results that deviated from the ground truth. Models trained with different modalities capture distinct feature distributions and characteristics such as textures and geometries. We assume the output deviation between the modalities is very different from each other. If the network output is reliable, then the outputs between modalities are well-aligned. This can be measured by the percentage overlap between output bounding boxes as well as the segmentation masks. Based on this assumption, we argue that the more deviations between the modalities there are, the less certain the output is. To estimate the deviation, we use Intersection Over Union (IOU). It is a ratio between the intersection of the two modalities and their union and can be applied to boxes as well as masks. Given a pair of detection/segmentation output x_0 and x_1 . Each represents a set of pixels. x_0 and x_1 can either be a pair of boxes or a pair of masks. Then, IOU can be calculated by:

$$IOU(x_0, x_1) = \frac{|x_0 \cap x_1|}{|x_0 \cup x_1|} \quad (4)$$

Where $|\cdot|$ is a function that computes the number of pixels for the given input. When deviation becomes larger, IOU will be smaller. When there is less deviation, IOU will be larger and close to 1. This behavior captures well the output alignment between modalities. When comparing results in object detection/segmentation, object matching is involved. There can be multiple detections for one object, so we average the IOU score for all related detections associated with this object. For simple annotation, we call the two models being compared source and target. Source results are matched to the target results. Let the set of all n_s detections in source be $D_s = \{d_{s,l} | l \in [0, n_s)\}$. We compute the IOU of all items in D_s associated with the k th target detection $d_{t,k}$, and obtain a set of IOUs $I_{D_s, d_{t,k}} = \{IOU(d_{s,i}, d_{t,k}) | d_{s,i} \in D_s\}$. We define objects with IOU lower than 30%, a typical threshold value used in the Non-Maximum Suppression (NMS) step in object detection networks such as [13], as non-matched and remove them. The updated IOU set is

$$I'_{D_s, d_{t,k}} = \{a | a \in I_{D_s, d_{t,k}}, a > 0.3\} \quad (5)$$

Next, we compute the average IOU $A(D_s, d_{t,k})$ for source detections D_s and target detection $d_{t,k}$:

$$A(D_s, d_{t,k}) = \text{mean}(I'_{D_s, d_{t,k}}) \quad (6)$$

We further compute the mean IOU for all n_t detections in target D_t :

$$mIOU(D_s, D_t) = \text{mean}(\{A(D_s, d_{t,k}) | k \in [0, n_t)\}) \quad (7)$$

We extend this mIOU score to describe the alignment between our network output and all the modalities. We name this score multimodal consistency (MC) score. MC can be used to describe the alignment of one modality or multiple modalities in a multimodal system. Let D_o to be the network output with n_o number of detections, and D_m to be the network output using only modality m . The MC score S_m for a single modality m is calculated by:

$$S_m = mIOU(D_m, D_o) \quad (8)$$

The MC score S for all modalities is computed by:

$$S = \text{mean}(\{A(D_m, d_{o,k}) | m \in [0, N), k \in [0, n_o)\}) \quad (9)$$

The higher the MC score is, the more reliable the system is. A score of 100% means all modalities predict the exact same output, and the system is very reliable. A score of 0% means each modality predicts a different output, and the system is unreliable.

4 Experiments

4.1 Dataset and Implementation Details

Among robotic grasping datasets, the MetaGraspNet dataset [11] provides large-scale, high-resolution simulated RGB and depth data as well as real-world data from an industry-grade sensor system. In addition, the dataset contains 82 objects and has a novel object set for testing. We divide the real dataset into train, validation, test, and test novel. We first exclude all scenes with novel objects, adding them to a separate novel test data split. Then we split the rest of the real dataset into 80% train, 10% validation, and 10% test.

Network	Box mAP	Mask mAP	Class agnostic		Class agnostic novel	
			Box AP	Mask AP	Box AP	Mask AP
Baseline-RGB	80.6	79.3	83.6	78.9	29.7	32.7
Baseline-Depth	73.1	71.9	77.2	72.0	32.3	32.1
Baseline-Fusion[1]	80.3	79.4	82.8	78.8	28.7	33.1
MSG Fusion	81.7	80.7	84.5	80.3	30.8	35.0

Table 1: Modality feature fusion comparison. Results (in %) are compared to Mask R-CNN [13] unimodal baselines as well as multimodal feature fusion approaches.

Network	Box		Mask		RGB off		Depth off	
	AP	MC	AP	MC	Box AP	Mask AP	Box AP	Mask AP
Baseline-Fusion[1]	82.8	43.0	78.8	41.0	0	0	75.2	68.6
MSG Fusion	84.5	39.1	80.3	36.7	0.6	0.3	78.1	74.3
+ dynamic ensemble	84.9	82.9	80.2	84.7	76.6	69.6	84.0	79.3

Table 2: Class-agnostic results for multimodal architectures illustrating the need for dynamic ensemble learning for multimodal redundancy. Best results highlighted in **bold**.

Due to the unique characteristics of each modality, we normalize and pre-process RGB inputs and depth inputs differently. We use the standard mean variance normalization for RGB inputs and we apply min-max normalization per scene for depth inputs, where depth values are min-max normalized to $[0, 1]$. We further flip the depth values to make 0 as the depth of the background and 1 as the closes point to the camera. With this value flip, background values are aligned to be 0 in each scene. In addition, this normalization added a data augmentation to the dataset as it stretches and compress object shapes in depth, allowing a fully utilized depth range where every depth value is used by an object.

Near objects’ edges, reflective surfaces, and transparent surfaces, there are often undefined values caused by a lack of signal returning to the depth sensor. As a pre-processing step, we apply image inpainting [2] to the depth images to replace any invalid values.

We use a classic object detection and segmentation network Mask-RCNN [13] with ResNet50 [12] backbone as our baseline. All the networks in our experiment are initialized by the same ImageNet [25] pretrained weights. We train all models with the same training configuration in terms of batch size, training epoch, and optimizer. We pretrain all the models on the simulated dataset of MetaGraspNet, and finetune on the real dataset. We report the performance of our method on the real test set with bounding box mean average precision (box mAP) and segmentation mask mean average precision (mask mAP).

4.2 Results and Discussions

Multi-scale Soft-Gate Feature Fusion (MSG Fusion): In Table 1, we show the improvement using our MSG Fusion under several train/test condition (regular class, class-agnostic, and novel objects). In general, we have observed that RGB data serves as a superior input modality for known objects. However, the performance of depth data tends to catch up in the detection of novel objects, as depth is more focused on object geometry. RGB image data in general provides rich texture features that can be easily used to differentiate items, but may become a distraction for the network leading to false positives. On the other hand, depth data provides simple geometric information that depicts object’s shape with smooth surfaces. Depth modality provides better performance when there is no RGB texture, dense texture creates distractions, and objects are too dark. The performance improves even more when RGB and depth are both utilized. The results show that dense fusion module provides better performance for novel items, while gated fusion strategy still provides competitive results. The slight performance drop from dense fusion to gating mechanism can be due to the simplicity of the network without the gating module. The gating module limits the information flow, but enables the independent training of sub-systems.

Data Split	MC Score - Box	MC Score - Mask
train	91.4	92.7
test	82.9	84.7
test-novel	73.4	73.7

Table 3: Class-agnostic MC scores from different data splits

Dynamic Ensemble Learning Strategy for Multimodal System: In this ablation study, we train our networks with and without the dynamic ensemble learning strategy and evaluate the networks’ performance under three different input conditions: both RGB and depth, depth-only, and RGB-only. As shown in Table 2, the Gate Fusion network performs well after both standard learning and dynamic ensemble learning when both inputs are present. While dynamic ensemble learning yields slightly lower mask AP (-0.1 AP), it also results in greater box AP (+0.4 AP). More notably, when the RGB input is removed, the network trained without the dynamic ensemble strategy fails catastrophically (i.e., -83.9 box AP, -80.0 mask AP). This indicates that the network is over-reliant on RGB image information due to that fact that RGB images were always available during training. In contrast, the network trained with dynamic ensemble learning is able to mitigate the loss of RGB information by leveraging depth information alone. While the overall performance is markedly reduced in this scenario (as expected based on the Baseline-Depth results in Table 1), it is a drastic improvement over the baseline Gate Fusion model. When depth inputs are turned off, the baseline model is still capable of reasonable operation but experiences a marked drop in performance. Adding dynamic ensemble learning mitigates this and results in only a minor loss of performance.

Interpreting Gates: The use of multi-scale soft gating offers a way to interpret the dynamic behaviour of a model under different input conditions. By averaging gate values over channels for a particular input scenario, we can obtain a human-interpretable heatmap for each input modality which illustrates the primary spatial regions being selected for by the gate. Figure 1 illustrates this technique for the soft gates at the second-highest spatial resolution of the feature pyramid. When both inputs are present, the soft gating mechanism primarily focuses on RGB features, although depth features are weighted strongly around the edges of the fruits (likely due to their homogeneity in RGB). However, once the RGB input is removed, the weights switch over to depth features dramatically in an attempt to compensate for the missing RGB features. Similarly, when depth features are removed, we see that RGB features are weighted more heavily around the edges of the fruits. This dynamic weight shifting between modalities is precisely the purpose of the soft gating mechanism and dynamic ensemble training strategy.

MC Score: The model exhibits a steady decline in MC score between the training set, test set, and test-novel set as seen in Table 3. This shows that the MC score decreases accordingly the more out-of-distribution a dataset is, correlating well with theoretically how reliable the model will be on each dataset. The MC score also differs dramatically between objects of different classes. Objects with poor MC scores include disinfection bottle, glass bottle, cables in transparent bag, eyeglasses, and so forth, while boxes, cups, cables (not in plastic bags), and pears have higher MC scores, as seen in Table 4. This shows the ability of the MC score to identify challenging objects in the dataset. We also compute the MC score against only RGB input or only depth input. Some objects exhibit a significant difference in MC score between those two options in Table 5. The starkest contrast appears when the object’s material properties result in significant noise in one sensor, such as when transparent or reflective objects causes errors in the depth sensor. Through this, we can identify when the model is highly reliant on a particular sensor for its predictions. Eyeglasses, for example, performs both poorly overall (Table 4, Figure 5d), and relies heavily on RGB input due to the transparency and reflection of its glass component.

Examples of object-level MC scores for segmentation mask detections can be seen in in Figure 5. Note that the confidence score predictions for each object remains at an inflated 0.999 for all four examples, while the MC score shows a greater distinction between the objects depending on difficulty. This is especially true in Figure 5d, where the model outputs a poor detection, but with high confidence. This, supported by previous dataset-level results, shows that the MC score is a better indicator for model reliability and uncertainty compared to the confidence score.

		MC Score (%)	
		Box	Mask
non-novel objects	disinfection bottle	60.5	62.6
	glass bottle	76.9	64.7
	cups d	96.0	96.5
novel objects	cables in transparent bag	68.4	61.0
	eyeglasses	73.6	63.7
	pear	77.6	91.4

Table 4: MC scores for different object classes

Class	Mask MC Score (%)	
	RGB only	Depth only
power drill	68.4	73.3
wineglass	89.8	76.5
eyeglasses	69.6	58.0

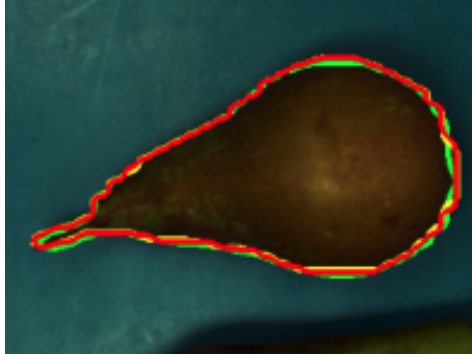
Table 5: Comparing MC score using only RGB or depth



(a) Train set object
MC score: 0.964
Confidence score: 0.999



(b) Test set object
MC score: 0.778
Confidence score: 0.999



(c) Test-novel set object
MC score: 0.966
Confidence score: 0.999



(d) Test-novel set object
MC score: 0.633
Confidence score: 0.999

Figure 5: Examples of object level MC scores. Gate fusion output is marked in red contour. RGB and depth only are marked in yellow and green contours respectively. In Figure 5b, some predictions identified the object as separate boxes, decreasing the MC score.

5 Conclusion

In this paper, we addressed the reliability aspect of deep learning-based computer vision systems for robotic grasping by introducing a multimodal redundancy framework called MMRNet. In particular, we realized the multimodal redundancy by two architecture changes, including a multi-scale dense modality feature fusion module and a gate fusion module to allow simple modality selection at per feature level. We also introduced a dynamic ensemble learning strategy to train modality models independently and collaboratively at the same time. Finally, we proposed a multimodal consistency score as a certainty indicator for the network output. Results show that our MMRNet provides reliable and superior performance in the event of a modality input failure. We could also show that MC score is an well suited output reliability indicator independent of the network’s confidence score.

Acknowledgments and Disclosure of Funding

We would like to thank the German Federal Ministry of Economic Affairs, the National Research Council Canada, Festo, and DarwinAI for supporting this work.

References

- [1] S. Back, J. Lee, T. Kim, S. Noh, R. Kang, S. Bak, and K. Lee. Unseen object amodal instance segmentation via hierarchical occlusion modeling. *arXiv preprint arXiv:2109.11103*, 2021.
- [2] M. Bertalmio, A.L. Bertozzi, and G. Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–I, 2001. doi: 10.1109/CVPR.2001.990497.
- [3] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
- [4] H. Cao, H. Fang, W. Liu, and C. Lu. Suctionnet-1billion: A large-scale benchmark for suction grasping. *IEEE Robot. Automat. Lett. (RA-L)*, 6(4):8718–8725, 2021.
- [5] J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7088–7097, October 2021.
- [6] V. Chordia and V. Kumar. Large scale multimodal classification using an ensemble of transformer models and co-attention. In *Proceedings of the SIGIR Workshop On eCommerce*, 2020.
- [7] F. F. dos Santos, P. Navaux, L. Carro, and P. Rech. Impact of reduced precision in the reliability of deep neural networks for object detection. In *IEEE European Test Symposium*, pages 1–6, 2019. doi: 10.1109/ETS.2019.8791554.
- [8] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Gläser, F. Timm, W. Wiesbeck, and K. Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2021. doi: 10.1109/TITS.2020.2972974.
- [9] Y. Freund, R. E Schapire, et al. Experiments with a new boosting algorithm. In *Proceedings of the International Conference on Machine Learning*, volume 96, pages 148–156, 1996.
- [10] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.
- [11] M. Gilles, Y. Chen, T. R. Winter, E. Z. Zeng, and A. Wong. MetaGraspNet: A large-scale benchmark dataset for scene-aware ambidextrous bin picking via physics-based metaverse synthesis. In *Proceedings of the International Conference on Automation Science and Engineering*, 2022.
- [12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, June 2016. Las Vegas, NV.

- [13] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, October 2017. doi: 10.1109/iccv.2017.322. Venice, Italy.
- [14] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang. What makes multi-modal learning better than single (provably). In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10944–10956, 2021. URL <https://proceedings.neurips.cc/paper/2021/file/5aa3405a3f865c10f420a4a7b55cbff3-Paper.pdf>.
- [15] Jeffrey Ichnowski, Yahav Avigal, Yi Liu, and Ken Goldberg. Gomp-fit: Grasp-optimized motion planning for fast inertial transport. *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 5255–5261, 2022.
- [16] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning*, pages 448–456, 2015. URL <http://jmlr.org/proceedings/papers/v37/ioffe15.pdf>.
- [17] G. Joshi, R. Walambe, and K. Kotecha. A review on explainability in multimodal deep neural nets. *IEEE Access*, 9:59800–59821, 2021. doi: 10.1109/ACCESS.2021.3070212.
- [18] S. Kumra and C. Kanan. Robotic grasp detection using deep convolutional neural networks. In *Proceedings of the International Conference on Intelligent Robots and Systems*, pages 769–776, 2017. doi: 10.1109/IROS.2017.8202237.
- [19] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. doi: 10.1109/CVPR.2017.106. Honolulu, HI.
- [20] A. Marinoni, S. Chlaily, E. Khachatrian, T. Eltoft, S. Selvakumaran, M. Girolami, and C. Jutten. Enhancing ensemble learning and transfer learning in multimodal data analysis by adaptive dimensionality reduction. *arXiv preprint arXiv:2105.03682*, 2021.
- [21] S. S. Menon and K. Krishnamurthy. Multimodal ensemble deep learning to predict disruptive behavior disorders in children. *Frontiers in Neuroinformatics*, 15, 2021. ISSN 1662-5196. doi: 10.3389/fninf.2021.742807.
- [22] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the International Conference on Machine Learning*, pages 689–696, June 2011. URL https://icml.cc/2011/papers/399_icmlpaper.pdf.
- [23] D. Ramachandram and G. W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017. doi: 10.1109/MSP.2017.2738401.
- [24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640*, May 2016. URL <http://arxiv.org/abs/1506.02640>.
- [25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, December 2015.
- [26] P. Santhanam, E. Farchi, and V. Pankratius. Engineering reliable deep learning systems. *arXiv preprint arXiv:1910.12582*, 2019.
- [27] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, 2014. URL <https://proceedings.neurips.cc/paper/2014/file/00ec53c4682d36f5c4359f4ae7bd7ba1-Paper.pdf>.
- [28] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004. doi: 10.1023/B:STCO.0000.

- [29] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang. Deep multimodal fusion by channel exchanging. *Proceedings of the Conference on Neural Information Processing Systems*, December 2020.
- [30] David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. ISSN 0893-6080. doi: 10.1016/S0893-6080(05)80023-1.
- [31] H. Xu, J. Blanchet, M. P. Gerardo-Castro, and S. Paudel. Measuring reliability of object detection algorithms for automated driving perception tasks. In *Proceedings of the Winter Simulation Conference*, pages 1–12, 2021. doi: 10.1109/WSC52266.2021.9715295.
- [32] Z. Xue and R. Marculescu. Dynamic multimodal fusion. *arXiv preprint arXiv:2204.00102*, 2022.
- [33] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018.
- [34] Q. Zhou, H. Liang, Z. Lin, and K. Xu. Multimodal feature fusion for video advertisements tagging via stacking ensemble. *arXiv preprint arXiv:2108.00679*, 2021.
- [35] Z. Zhou. *Machine Learning*. Springer Singapore, 08 2021. ISBN 9789811519673.

A Additional Results

We show complete Multimodal Consistency (MC) score results for our real-world test-novel experiment in Table 6. This table contains results for all object categories in the test-novel set which was unable to fit into the page limit. We also show randomly selected image results from this experiment for every object categories in test-novel dataset in Figure 6.

In addition, we show randomly selected image results on the real-world test dataset in Figure 7.

Table 6: MC scores (%) for all object categories in the test-novel set. Item order is ranked based on the combined MC score for mask from low to high.

Class	Fig 6 ID	MC Score (%) - Mask			MC Score (%) - Box		
		combined	image	depth	combined	image	depth
cables in							
transparent bag	(a)	61.0	70.7	50.8	68.4	76.6	75.1
eyeglasses	(b)	63.7	69.6	58.0	73.6	81.1	79.0
power drill	(c)	63.9	68.5	73.4	68.0	77.8	74.6
clamp	(d)	68.6	67.0	74.3	75.2	83.6	80.9
marker small	(e)	77.8	83.9	72.6	70.8	80.3	78.4
mug	(f)	78.6	87.1	69.2	71.6	86.6	82.8
marker big	(g)	79.2	86.9	84.3	67.4	77.9	77.1
clamp big	(h)	79.4	83.6	78.2	74.6	82.7	80.7
clamp small	(i)	79.7	82.7	80.0	75.5	84.0	81.1
wineglass	(j)	79.7	89.8	76.6	72.7	81.2	79.5
crayons	(k)	81.1	83.6	78.9	71.7	79.8	77.7
cables	(l)	85.3	89.3	86.1	77.1	88.5	87.8
pear	(m)	91.4	94.3	93.0	77.6	88.8	84.7

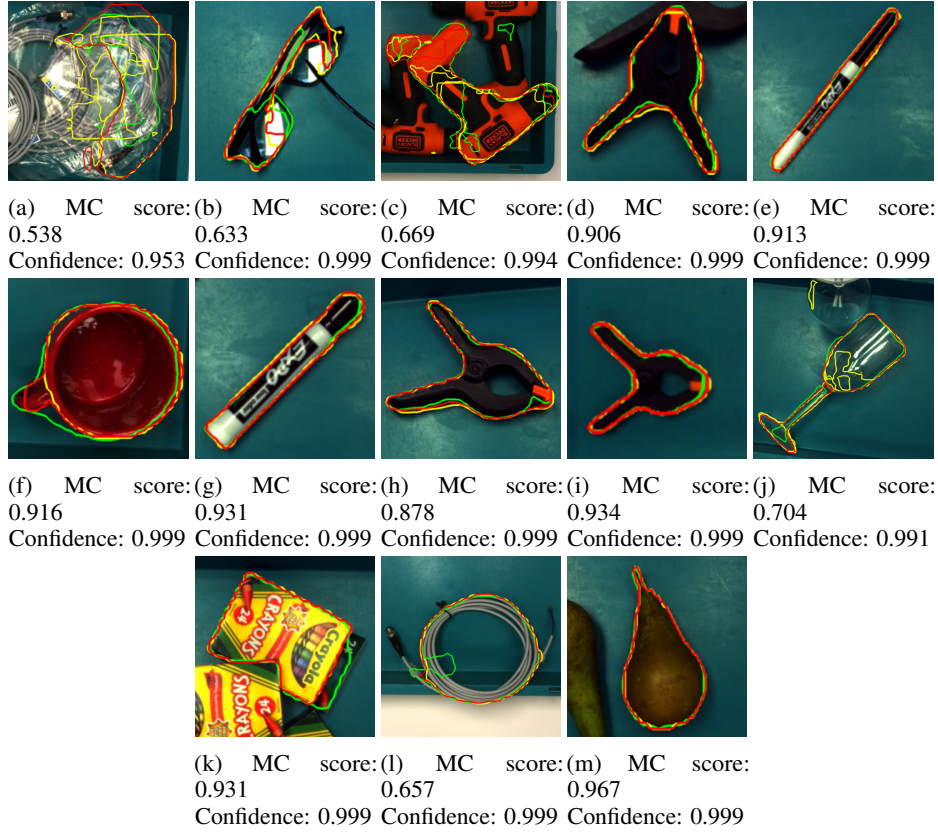


Figure 6: Image examples of objects from the novel-test set. The MC score shown is calculated at per object level instead of being averaged over the entire object category. We also show output confidence score for comparison. Gate fusion output is marked in red contour. RGB and depth only are marked in yellow and green contours respectively.



Figure 7: Image examples of objects from the real-world test dataset. The MC score shown is calculated at per object level instead of being averaged over the entire object category. We also show output confidence score for comparison. Gate fusion output is marked in red contour. RGB and depth only are marked in yellow and green contours respectively.