Modality-Incremental Learning with Disjoint Relevance Mapping Networks for Image-based Semantic Segmentation

Niharika Hegde [1,2] * Shishir Muralidhara [2] * René Schuster [1,2] Didier Stricker [1,2]

[1] RPTU – University of Kaiserslautern-Landau

[2] DFKI – German Research Center for Artificial Intelligence

firstname.lastname@dfki.de

Abstract

In autonomous driving, environment perception has significantly advanced with the utilization of deep learning techniques for diverse sensors such as cameras, depth sensors, or infrared sensors. The diversity in the sensor stack increases the safety and contributes to robustness against adverse weather and lighting conditions. However, the variance in data acquired from different sensors poses challenges. In the context of continual learning (CL), incremental learning is especially challenging for considerably large domain shifts, e.g. different sensor modalities. This amplifies the problem of catastrophic forgetting. To address this issue, we formulate the concept of modality-incremental learning *and examine its necessity, by contrasting it with existing incremental learning paradigms. We propose the use of a modified Relevance Mapping Network (RMN) to incrementally learn new modalities while preserving performance on previously learned modalities, in which relevance maps are disjoint. Experimental results demonstrate that the prevention of shared connections in this approach helps alleviate the problem of forgetting within the constraints of a strict continual learning framework.*

1. Introduction

Continual learning (CL) has emerged as a fundamental paradigm to address the need for intelligent agents to continually update with new information while preserving learned knowledge. In contrast, conventional machine learning normally builds on a closed dataset, *i.e.* it can only handle a fixed number of predefined classes or domains, and all the data needs to be presented to the model in a single training step. However, in practical scenarios, models frequently face the challenge of dealing with changing data and objectives. This problem can be circumvented by accumulating all data and retraining the model to derive a unified model effective across a combined dataset. Although this approach achieves optimal performance, it is often impractical and may not be feasible due to several reasons. For instance, anticipating future data is not possible in real-world applications, and access to previous data might be restricted due to privacy concerns or resource constraints. Moreover, retraining from scratch using all past data results in a significant increase in training time and computational requirements. Consequently, learning solely from new data is more efficient, but can lead to catastrophic forgetting [29], where past knowledge is overwritten resulting in degraded performance on the previous tasks. This challenge emphasizes the importance of developing CL methods to maintain a balance between incorporating new information and retaining past knowledge, referred to as the stability-plasticity dilemma [30].

Autonomous driving systems are typically trained on normal driving conditions due to their prevalence and ease of accessibility. However, as these systems advance, they must confront a multitude of driving scenarios, including adverse weather, low-light conditions, and other challenging environments. This shift in data distribution, can undermine their ability to make precise predictions or decisions, raising potential safety concerns. Single sensor systems, in particular, struggle to adapt to challenging conditions which can severely impact their performance. Integrating a multimodal, complementary sensor suite is an effective measure to encounter deficiencies under such changes of conditions. For example, IR cameras are effective under low-light conditions but can be affected by weather conditions like rain and fog. Depth sensors offer precise distance measurements but may be limited in range. Combining diverse sensors in a heterogeneous stack helps alleviate the limitations of individual sensor types and enhances the overall performance and reliability of autonomous systems.

* These authors have contributed equally to this work.

For an existing system, new sensor modalities might be introduced as they undergo technical advancements, become more cost efficient, or address specific limitations. In such cases, it's appealing to have a single, unified model that incrementally learns to handle the new modalities and enhances its ability to perceive under challenging driving conditions and varying sensor characteristics, without forgetting previously acquired knowledge. In this paper, we introduce and formalize this novel incremental setting termed *modality-incremental learning* (MIL) to learn on an extending set of sensor modalities and contrast it against existing incremental paradigms. We exemplify the concept of MIL by semantic segmentation on various visual modalities (*i.e.* RGB, IR, and depth cameras) in an automotive setting.

Current incremental settings typically use data from a single visual modality, and the methods designed for them lack the capability to manage changing modalities. Addressing this challenge of learning visual modalities, we propose the use of Disjoint Relevance Mapping Networks (DRMNs), which aim to learn an improved representational map, such that the significantly distinct tasks (changing modalities) use different subsets of the network parameters. We argue that the prevention of overlap in the relevance maps mitigates forgetting completely, without having a negative impact on the utilized network's capacity. The contribution of our work can be summarized as follows:

- We introduce and formulate the problem of modality-incremental learning (MIL) in the context of continual learning, and demonstrate it for semantic segmentation in an automotive context.
- We benchmark existing methods for domain-incremental learning (DIL) in this novel setting.
- We propose a modified version of Relevance Mapping Networks (RMN) [25] that is tailored towards MIL.
- We evaluate the proposed Disjoint Relevance Mapping Networks (DRMN) in terms of accuracy, forgetting, and network utilization on various MIL settings across two multi-modal datasets.

2. Related Work

Continual learning strategies can be categorized into three types: Architecture-based, replay, and regularization methods. Architecture-based methods address forgetting by altering the architecture of networks either explicitly or implicitly to learn new tasks. Explicit modification involves dynamically expanding the network architecture by adding individual neurons [44], widening/deepening layers [41], or cloning the network [35]. Implicit modifications use a fixed network capacity and adapt to new tasks through freezing [24], pruning [28] or task-specific paths [11]. Architecture-based methods also include dual-architecture models inspired by the brain [15, 26].

Replay-based methods address forgetting by replaying previously encountered information. These methods can be classified into experience replay and generative replay. Experience replay [17, 21] or rehearsal, involves storing a subset of instances from the previous task, which are later used during retraining on a new task. However, experience replay faces challenges related to privacy and storage of data. Generative replay [36, 42] methods diverge from rehearsal approaches by training generative models, allowing them to generate samples from previous tasks.

Regularization is a process of introducing an additional term into the loss function to regulate the update of weights when learning in order to retain previous knowledge. Regularization includes identifying crucial weights [1, 27, 45] within a model and preventing overwriting them, or storing learned patterns to guide the gradients [20, 23]. Distillation methods [13, 31] transfers knowledge from one neural network to another. Such methods do not need to store data, and only require a previous model for knowledge transfer.

In this work, we propose a hybrid approach that builds on RMNs [25] and combines architectural and regularization techniques. The idea is to maintain a fixed network capacity by freezing task-specific weights and utilize pruning to free weights for subsequent tasks. The relevance maps help in identifying the important weights from previous tasks, and we enforce parameter isolation by masking these weights.

2.1. Continual Semantic Segmentation

Continual semantic segmentation (CSS) constitutes a specialized sub-field within the broader realm of continual learning, focusing specifically on semantic segmentation. Most research in CSS follows either one out of two popular incremental learning schemes. The first is class-incremental learning (CIL) [3, 4, 10, 16, 46], in which sets of classes are learned sequentially. The second is domain-incremental learning (DIL), which is closer to the proposed MIL setting. Here, the distribution of input data is extended over time. In fact, MIL can be viewed as a severe form of DIL, in which individual sensor modalities represent entirely different visual domains. For domain-incremental semantic segmentation, MDIL [14] partitions the encoder network into domain-agnostic and domain-specific components to learn new domain-specific information, and a dedicated decoder is instantiated for each domain. DoSe [33] uses domain-aware distillation on batch normalization for incremental learning using a pretrained model. It also uses rehearsal for storing and replaying difficult instances from previously seen domains. Addressing the storage constraints in rehearsal-based approaches, Deng and Xiang [9] propose a style replay method to reduce storage overhead.

Our work is in contrast with the existing work by Barbato *et al*. [2] who use multiple modalities in a continual learning setting within the context of CIL. *I.e*., all modalities are used in all tasks. Their work assumes a pre-defined number of modalities, allowing for the design of suitable architectures. MIL in this work aligns more closely with DIL since the number of classes remains consistent across tasks.

2.2. Multi-Modal Semantic Segmentation

Early multi-modal segmentation methods [7] combined data from different modalities and used this combined input for the segmentation network. However, this strategy of early fusion struggles to effectively capture the diverse information provided by different modalities. Recent advancements aim to leverage the strengths of various modalities by employing multiple fusion operations at various stages of the network [18]. A common architectural choice involves a multi-stream encoder [8], where each modality has its own network branch. Additional network modules [22] connect these branches to combine modality-specific features across branches, facilitating hierarchical fusion.

For multi-modal segmentation using RGB and depth modalities, AsymFusion [40] uses a bidirectional fusion scheme with shared-weight branches and asymmetric fusion blocks to enhance feature interactions. Chen *et al*. [6] proposed a unified cross-modality guided encoder with a separation-and-aggregation gate (SA-Gate) for effective feature re-calibration and aggregation across modalities Mid-fusion architecture [32] combines sensor modalities at the feature level using skip connections for autonomous driving. CMX [47] leverages cross-modal feature rectification and fusion modules, integrating a cross-attention mechanism for enhanced feature fusion across modalities.

For multi-modal segmentation using RGB and IR modalities, ABMDRNet [48], uses a bi-directional image-to-image translation to mitigate modality differences between RGB and thermal features. GMNet [49] integrates multi-layer features using densely connected structures and residual modules, with a multistream decoder that decouples semantic prediction into foreground, background, and boundary maps. RTFNet [37] characterized by the asymmetrical encoder and decoder modules, merges modalities at multiple levels of the RGB branch. FuseSeg [38] proposed the hierarchical addition of thermal feature maps to RGB feature maps in a two-stage fusion process. CCAFFMNet [43] leverages multi-level channel-coordinate attention feature-fusion blocks within a coarse-to-fine U-Net architecture.

This work addresses multi-modal segmentation from a continual learning perspective, where modalities are incrementally and arbitrarily added. This complicates the design of specialized architectures for handling multiple modalities. Therefore, we process each modality independently for segmentation, leaving more advanced fusion techniques to the possibilities for future research.

Figure 1. Three different modalities to perceive traffic scenarios in an automotive context. From left to right: Classical RGB, depth, and IR images from the InfraParis dataset [12].

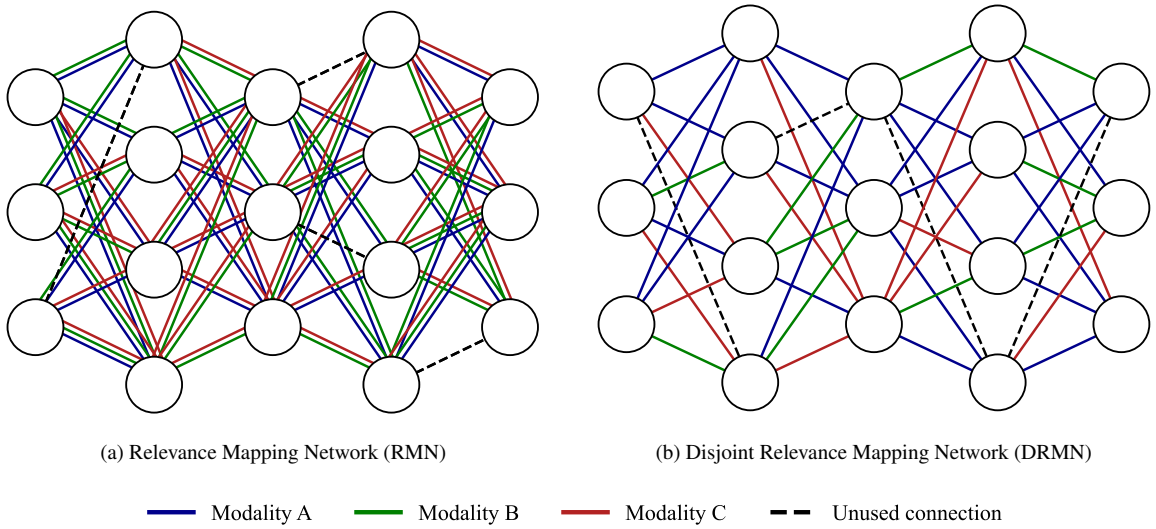
3. Modality-Incremental Learning (MIL)

Incremental learning involves learning a sequence of tasks $T = T_0, T_1, ..., T_n$. Each task T_i is associated with task-specific data $D_i = (X_i, Y_i)$, and represents a change either in the input or the output distribution. In domain-incremental learning (DIL), the input distribution X changes at each task increment, while the output distribution remains the same. Each task can represent different data sources such as geographical locations or weather conditions. In class-incremental learning (CIL), the input data remains constant, while each task introduces a subset of new classes C_i, such that $C_0 \cup C_1 \cup C_i = C \in Y$ the model has to learn without forgetting previously learned classes.

We introduce modality-incremental learning (MIL), a novel incremental learning setting tailored to handle the case of incrementally learned sensor modalities. In MIL, each new task with associated data (M_i, Y) presents a change in the input distribution by introducing a new modality M_i. The set of classes Y remains consistent across all tasks, similar to DIL. Unlike DIL, where the input X_i remains within the same visual modality across all tasks, MIL faces more significant data drift due to the introduction of new modalities. As a result, DIL methods struggle to adapt effectively to MIL scenarios, as shown in our experiments.

To underline this difference and the severity of the domain gaps between modalities, we highlight that even an offline training on joint data from all MIL tasks produces subpar results compared to modality-specific modules. In CL, this *joint training* usually forms a theoretical upper bound, since diverse data facilitates the learning and forgetting does not occur. However in MIL, the substantial differences between modalities pose a significant challenge for joint training to effectively utilize shared knowledge across tasks.

A notable advantage of MIL, compared to DIL or CIL, is the straightforward availability of the task ID, as it can be safely assumed that the sensor that produces the input signal is known to the system. This inherent knowledge obviates the need for explicit task identification during inference.



(a) Relevance Mapping Network (RMN)

(b) Disjoint Relevance Mapping Network (DRMN)

Figure 2. Relevance Mapping Network (RMN) (left) shares connections across tasks, with new tasks utilizing their respective relevance map values and the previous weights. In contrast, the Disjoint RMN (DRMN) (right) isolates connections between tasks, compelling the network to learn independent, task-specific weights and mitigates the negative interference when incrementally learning modalities. It is important to note that each node can be used for a modality-specific representation in all tasks.

4. Disjoint Relevance Mapping Networks

The challenge of learning multiple modalities lies in the inability of a single encoder to manage them all, even in an offline setting, and this is exacerbated when learning modalities incrementally. This limitation renders most continual learning methods, such as distillation and rehearsal or replay, ineffective as they still rely on a single network. Architecture-based methods, such as multi-encoder or multiple networks show promise but do not scale well with an increasing number of tasks. The number of models and storage requirements grow proportionally with each new modality. In light of these limitations, it would be desirable to have a single model of fixed size that can effectively handle various modalities, unlike previous methods. To this end, we propose using Relevance Mapping Networks (RMNs) [25] to handle incremental learning of modalities. This approach requires the task ID to be known during inference, which is not an issue in MIL as explained earlier. We further modify the original RMN concept with parameter isolation to better fit the needs for MIL.

4.1. Relevance Mapping Networks

RMNs are a method inspired by the optimal overlap hypothesis, which aim to learn an optimal representational overlap, such that unrelated tasks use different network parameters, while allowing similar tasks to have a representational overlap. RMN was originally proposed for image classification in the continual learning setting [25]. In this work, we extend the implementation of RMN beyond image classification, to tackle the complex task of continual semantic segmentation. RMNs enhance existing neural networks by augmenting the convolutional and linear layers with additional weights referred to as relevance maps $\mathbb{M}$ as illustrated in Fig. 3. These relevance maps are unique to each task and identify the most important neural connections within the network for the corresponding task, and are used in conjunction with the standard layer weights $\mathbf{W}$:

$$f_{out} = \mathbf{W} \cdot \mathbb{M}_t \cdot f_{in} \tag{1}$$

The learned relevance maps can be interpreted as (soft) masks, selecting and freezing the crucial task-specific weights in the network, resulting in dynamic task-specific paths. This approach effectively dissects the network into partial subnetworks, while still allowing it to share information across related tasks and maintaining task-specific weights.

The relevance maps $\mathbb{M}$ are learned continuously with a bounded activation in the interval $[0..1]$. During training, the RMN periodically applies thresholding with a hyperparameter μ, called the pruning parameter, to select relevant connections, of which the corresponding values in $\mathbf{W}$ are frozen, and to set irrelevant connections to zero (both in $\mathbb{M}_t$ and $\mathbf{W}$). This way, unused capacity of the network is freed for future tasks.

4.2. Disjoint Relevance Mapping Networks for MIL

A key challenge in using RMNs for modality-incremental learning is to balance the utilization of network capacity and the overlap between tasks. Especially in MIL,

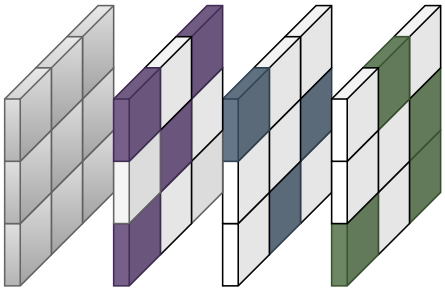


Figure 3. A Relevance Mapping Network augments the network weights by adding task-specific relevance maps $\mathbb{M}$ to select the important weights for each task.

we argue that too much overlap between the task-specific network paths hinders learning and amplifies forgetting, due to the naturally large differences between modalities (*cf*. Sec. 3). With regular RMNs, we observe a significant overlap of used connections across tasks. While the freezing of relevant weights helps retain knowledge from previous tasks, it lacks a mechanism to promote increased adaptation to drastically different sensing modalities. Too much overlap in relevance maps forces the networks to reuse the previously learned weights, and with highly disparate modalities, this leads to inadequate learning on the new task. We demonstrate under Sec. 5.4, the paradoxical effect of weight sharing, which typically is beneficial for knowledge transfer, but becomes detrimental in this context due to modality-specific conflicts.

To address this issue of overlap, we propose Disjoint Relevance Mapping Networks (DRMNs). As the name suggests, DRMNs enforce a complete separation of relevant neural connections between modalities. The idea and differences to classical RMNs are visualized in Fig. 2. RMN allows sharing connections across tasks, with new tasks potentially reusing the previously learned connections with the newly learned relevance maps. In contrast, DRMN uses parameter isolation, and the network is forced to learn task-specific connections for each new task. By doing so, DRMN aims to reduce interference and conflicts that can arise when learning diverse modalities. To enforce parameter isolation, each relevance map of the previous tasks $\mathbb{M}_i \forall i < t$ is analyzed to identify used connections. These values are then set to zero in $\mathbb{M}_t$, effectively rendering them unimportant for the current task. This limits the learning of $\mathbf{W}$ and $\mathbb{M}_t$ to connections not used in prior tasks. Intuitively, one might assume that this strict separation has a negative impact on the transfer of knowledge and the depletion of the network's capacity. However, our experiments in Secs. 5.3 and 5.5 show that this is not the case. While connections must not be shared across tasks, network nodes can be used in multiple modalities. This mechanism effectively minimizes negative interference between tasks during learning, but allows to learn powerful representations from sparse connections. In fact, forgetting is reduced due to the increased decoupling, while the overall utilization of connections is barely affected, compared to original RMNs. The incremental training process with DRMNs is detailed in Algorithm 1.

Algorithm 1 Disjoint Relevance Mapping Network

Training Phase
Input: Training data (X, Y), task IDs $t = 0, 1, \ldots, n$, prune parameter μ, initialized relevance maps $\mathbb{M}$
for $t = 0$ to n **do** ▷ Train on task t
  if $t > 0$ **then**
    $m_{\textbf{unused}} \leftarrow \bigwedge_{i=0}^{t-1}(\neg \mathbb{M}_i)$
  else
    $m_{\textbf{unused}} \leftarrow \mathbf{1}$
  end if
  $f(X_t; \mathbf{W}; \mathbb{M}_t) \implies \hat{Y}_t \leftarrow \sigma((\mathbf{W} \cdot \mathbb{M}_t \cdot m_{\textbf{unused}}) \cdot X_t)$
  Relevance Map Pruning, $\mathbb{M}_t \leq \mu$
  Freeze weights in f where $\mathbb{M}_t \neq 0$
end for
Inference Phase
Input: Task ID t is given by the sensor and used to select the relevance map $\mathbb{M}_t$ for that task.
Output: $f(X; \mathbf{W}, \mathbb{M}_t)$

5. Experiments

5.1. Datasets

To simulate a sequentially adapted sensor system, we utilize datasets that offer ground truth for our task, *i.e*. pixel-wise semantic labels, as well as input images captured with different sensors. The datasets Freiburg Thermal [39] and InfraParis [12] offer aligned RGB and infrared images. The InfraParis dataset offers an additional visual modality in the form of depth maps, enhancing the diversity of available modalities. To further enhance the versatility of the experimental setup, an additional visual modality in the form of grayscale images was created for both the datasets. This diversity allows for a more comprehensive assessment of the proposed approaches for MIL, as well as facilitating the exploration of the effects of incrementally learned modalities.

- **Freiburg Thermal** [39] encompasses diverse driving scenes such as highways, cities, suburbs, and rural areas with pixel-level labels for 13 object categories. Including the simulated grayscale sensor, it covers 3 modalities. The dataset has 9,735 images for training and 2,435 for validation for each modality.

- **InfraParis** [12] dataset contains RGB, depth, and infrared data captured in various cities around Paris.

Table 1. Results for three MIL task sequences on Freiburg Thermal [39] dataset after learning all tasks.

Method	RGB→IR→Gray				Gray→RGB→IR				IR→Gray→RGB			
	RGB	IR	Gray	**Avg**	Gray	RGB	IR	**Avg**	IR	Gray	RGB	**Avg**
Single Task	76.41	59.56	74.56	*70.18*	74.56	76.41	59.56	*70.18*	59.56	74.56	76.41	*70.18*
Joint Training	75.43	56.06	74.46	*68.65*	74.46	75.43	56.06	*68.65*	56.06	74.46	75.43	*68.65*
Fine Tuning	74.97	07.41	74.88	*52.42*	13.76	14.06	60.19	*29.34*	07.12	73.17	75.24	*51.84*
EWC [27]	72.28	07.98	72.93	*51.06*	18.38	18.02	40.80	*25.73*	10.56	59.47	61.69	*43.91*
ILT [31]	74.02	07.91	66.27	*49.40*	13.44	13.40	08.96	*11.93*	20.68	22.57	23.53	*22.26*
RMN [25]	73.13	55.01	68.29	*65.48*	71.09	72.82	53.57	*65.83*	55.10	68.90	69.46	*64.49*
DRMN (Ours)	73.21	54.95	69.38	***65.85***	71.12	72.61	54.12	***65.95***	54.97	70.56	71.19	***65.57***

This yields 4 different sensor modalities, when augmented with grayscale images. The dataset offers pixel-wise annotations for 20 classes. It contains 3545 RGB images as well as 6567 Depth and IR images (each) for training and 189 images per modality for validation. For depth images, we mask ground truth labels wherever the corresponding depth value is zero.

5.2. Implementation and Baselines

We evaluate the results of our approach against existing continual learning approaches and the standard baselines, *i.e. joint training*, *fine-tuning*, and *single-task learning*. Joint training learns the set of tasks concurrently in a single step, using all modalities. Since the model is trained on all tasks simultaneously, there is no catastrophic forgetting and the learning benefits from the extended dataset size. The single-task baseline refers to a model trained on just one task without any further adaptation or incremental steps. In the context of MIL, the single-task models have been trained on individual modalities and serve as the upper bound for comparison. Fine-tuning is a naive solution for incremental learning, in which a model is trained sequentially for each task, building on the previously learned tasks. While effective for learning new tasks, it suffers most from forgetting of previous tasks, especially with modalities that are considerably different, *e.g.* RGB and IR images. In our experiments, we further compare against two regularization-based approaches, Elastic Weight Consolidation (EWC) [27] which penalizes overwriting of important weights, and Incremental Learning Techniques (ILT) [31] which uses knowledge distillation. We use the Relevance Mapping Network (RMN) from [25] adapted for segmentation as our baseline for comparison. The experiments highlight its shortcomings in handling multiple modalities and learning new modalities. Our proposed Disjoint RMN (DRMN) overcomes this limitation by ensuring each task learns independent weights.

For the segmentation network, we use a DeeplabV3+ [5] model with a ResNet-101 [19] backbone, which is pretrained on the ImageNet [34] dataset. This backbone is used for all initial, single and joint-training models. In incremental learning methods, the previous task model serves as the starting point. As part of our work, we use a custom DeepLabV3+ model that incorporates relevance maps into the convolutional layers, along with task-specific batch normalization [25]. All methods used for comparison are trained on a single RTXA6000 GPU with a batch size of 2 for 75 epochs, using the stochastic gradient descent (SGD) optimizer with a learning rate of 1e-5 across all tasks. The prune parameter μ is used as a threshold for determining the important weights in the relevance maps. We use $\mu = 0.6$ and weights below this value are pruned after each epoch starting from epoch 50. Results for different threshold values are provided in our supplementary material. Our experiments are evaluated in terms of mean Intersection-over-Union (mIoU).

5.3. Benchmark

5.3.1 Freiburg Thermal

Using the three modalities from the Freiburg Thermal dataset [39], we design the following non-exhaustive task sequences. These sequences cover all possible learning orders for each modality: (***RGB*** → ***IR*** → ***Gray***), (***Gray*** → ***RGB*** → ***IR***) and (***IR*** → ***Gray*** → ***RGB***).

From Tab. 1, we can observe that the best results are achieved using the single-task models. The joint training baseline falls short due to the higher complexity of learning different modalities. However, it does not suffer from catastrophic forgetting as all the modalities are learned in a single step. The fine-tuning approach exhibits positive forward transfer, surpassing even the single-task models on the final task. But this comes at the cost of significantly overwriting previous task information. It is also more sensitive to the task sequence, when the final modality learned is ei-

ther RGB or grayscale, then it benefits the other modality learned in the previous steps, as they share a higher degree of similarity. This trend can also be observed in EWC [27]. However, when highly dissimilar modalities are learned in the first and last step, the model fails to learn effectively, as it hinders and prevents overwriting of previous weights. ILT [31] uses knowledge distillation at both the feature and output levels. However, aligning features from different modalities can be detrimental, especially for task sequences where the initial and current modalities are vastly different. Both RMN [25] and our proposed DRMN are more robust and mostly unaffected by the order in the task sequences, achieving consistent results. Notably, across all three sequences, DRMN outperforms the baseline on the final task.

5.3.2 InfraParis

Using the four modalities from the InfraParis [12] dataset, we perform our experiments on the following sequence of tasks: (***IR*** $\rightarrow$ ***RGB*** $\rightarrow$ ***Depth*** $\rightarrow$ ***Gray***). Similar to the results from Freiburg Thermal [39] dataset, the single task models form the upper bound. The addition of another modality in the form of depth images underscores the challenges faced by a single model in handling diverse inputs, as observed in the joint training baseline. Task order influences fine-tuning results, with better results achieved when the last learned modality (such as gray) complements the previously learned modalities (RGB). However, due to the separation of task-specific weights, this positive transfer is lower in our DRMN. Regularization methods such as EWC [27] and ILT [31] highlight the difficulty of balancing between learning new tasks while retaining previous information. Both RMN [25] and our proposed DRMN achieve comparable results, surpassing the joint training baseline and nearly matching the performance of the single-task models.

5.4. Shared Weights in RMN for MIL

Relevance Mapping Networks (RMNs) preserve knowledge from prior tasks by freezing weights crucial for those tasks, which are identified using the relevance maps. However, this does not prevent the network from reusing the same weights for new tasks. It merely forces the network to reuse the old weight values with the newly learned relevance map, preventing any updates to the network's original weights. We hypothesize that in the context of MIL, where the input modalities are significantly different, sharing weights between tasks can be detrimental and potentially hinder learning. To validate this, we explored mechanisms to reduce weight overlap between tasks and force the network to learn independent weights. The results on Freiburg Thermal [39] are presented in Tab. 3. Initially, we tried to enforce this constraint through an additional loss term that calculates the overlap between the current task and previous tasks and penalizes it. The addition of overlap loss to RMN (ORMN) marginally reduces the overlap, which will be discussed further under Sec. 5.5, and achieves results similar to the baseline. This necessitates a more explicit approach to force the network to learn new weights, and we achieve this by masking the previous relevant weights during learning. We experimented with two variants of this approach: Partial masking with RMN (PRMN), which allows weight sharing in the decoder while preventing overlap in the encoder; and the proposed Disjoint RMN (DRMN), which completely prevents the weights from being shared. Both methods achieve better results compared to the baseline RMN. Notably, DRMN, with stricter weight separation, outperforms the more relaxed PRMN. This reinforces our hypothesis that weight sharing can be inhibiting in MIL, leading to conflicts when learning distinct modalities.

Table 2. Results on the InfraParis [12] dataset after learning all tasks.

Method	**IR→RGB→Depth→Gray**				
	IR	RGB	Depth	Gray	**Avg**
Single Task	38.11	62.01	55.25	61.59	*54.24*
Joint Training	31.86	61.65	34.64	61.18	*47.33*
Fine Tuning	24.84	60.67	16.19	58.83	*40.13*
EWC [27]	26.72	52.44	26.90	52.37	*39.61*
ILT [31]	35.79	28.38	24.62	27.85	*29.16*
RMN [25]	39.85	55.21	50.10	50.14	*48.82*
DRMN (Ours)	39.03	53.81	50.11	52.76	***48.92***

5.5. Task Overlap and Network Utilization

We previously highlighted the importance of learning independent and task-specific weights in Sec. 5.4 for MIL. In this section, we examine how different methods utilize the network and share weights across tasks of Freiburg Thermal [39], highlighting their influence on the results presented in Tab. 3. From Tab. 4, we can observe that RMN consistently has higher network utilization for each task, with nearly one-fourth of the weights shared between any two tasks and nearly half the weights shared across all tasks. ORMN, which uses overlap loss to deter weight sharing, exhibits a slight decrease in the percentage of shared weights. However, this marginal reduction has negligible impact on the performance of the incrementally learned modalities. PRMN, which utilizes partial masking, demonstrates a significant decrease in shared weights, leading to improved learning on new modalities. Using independent weights raises concerns about limited network capacity for future tasks. This concern is heightened with our disjoint RMN (DRMN) approach, which completely prevents weight shar-

Table 3. Results highlighting the influence of weight sharing for three MIL tasks on Freiburg Thermal [39] dataset. From top to bottom, the methods increase in how much separation of neural connections across tasks is enforced (*cf*. Sec. 5.4).

Method	RGB→IR→Gray				Gray→RGB→IR				IR→Gray→RGB			
	RGB	IR	Gray	**Avg**	Gray	RGB	IR	**Avg**	IR	Gray	RGB	**Avg**
RMN [25]	73.13	55.01	68.29	*65.48*	71.09	72.82	53.57	*65.83*	55.10	68.90	69.46	*64.49*
ORMN	73.07	54.75	67.66	*65.16*	71.06	72.53	52.79	*65.46*	55.11	69.02	69.00	*64.38*
PRMN	73.18	54.94	69.13	*65.75*	71.14	72.46	53.46	*65.69*	55.00	70.51	70.77	*65.43*
DRMN	73.21	54.95	69.38	***65.85***	71.12	72.61	54.12	***65.95***	54.97	70.56	71.19	***65.57***

Table 4. Analysis of task overlap and network utilization for task sequence (***IR*** → ***Gray*** → ***RGB***) using Freiburg Thermal [39] dataset.

Method	Task-wise Utilization			Pairwise Overlap and Utilization						Overall	
	IR	Gray	RGB	IR & Gray		IR & RGB		Gray & RGB		Overlap	Network
				Overlap	Util	Overlap	Util	Overlap	Util	Weights	Utilization
RMN [25]	47.79	47.78	47.80	22.85	72.73	22.84	72.76	22.85	72.73	46.68	85.77
ORMN	47.79	47.68	47.66	22.78	72.69	22.77	72.68	22.72	72.62	46.55	85.72
PRMN	47.79	24.95	16.11	0.00	72.74	2.02	61.88	1.06	40.00	3.08	85.77
DRMN	47.79	24.95	13.03	0.00	72.74	0.00	60.83	0.00	37.99	0.00	85.78

ing. However, Tab. 4 shows DRMN efficiently learns new tasks with similar network utilization despite having fewer available weights, alleviating concerns about network capacity exhaustion. An analog experiment on InfraParis [12] indicates the same trend. The exact utilization on that dataset is detailed in the supplementary material.

5.6. Efficiency and Scalability

Compared to growing architectures, DRMN maintains a constant size for any number of tasks. Duplicating encoders, decoders, or entire networks in each task introduces a significant linear growth of the overall model. Similarly, the original RMNs require one full relevance map for all weights for each task. However, the disjoint property of DRMNs allows for an efficient implementation that stores the relevance maps for all tasks in a single data structure.

The overhead of selecting and loading the appropriate relevance maps is also constant and negligibly small compared to the computation within the network.

6. Conclusion

This work introduces a new paradigm called modality-incremental learning (MIL). In contrast to existing incremental learning settings where the input distribution comes from the same visual modality, MIL addresses a larger domain gap between tasks, as the modalities can vary significantly. Consequently, existing continual learning methods and baselines fall short in handling multiple modalities, necessitating the development of tailored and dynamic approaches for MIL. We build upon the Relevance Mapping Network (RMN). Unlike dynamically growing architecture methods that raise scalability concerns, RMNs use a fixed network architecture and relevance maps to incrementally adapt to new tasks. We introduce a crucial modification with our Disjoint RMN (DRMN) by strictly separating neural connections across tasks. This approach demonstrates improved learning across modalities by reducing conflicts between them, though keeping the overall network utilization at a comparable level. For future work, we plan to implement an adaptive regularization term for the overlap between task-specific relevance maps that considers the similarity between modalities.

Acknowledgments

This work was partially funded by the Federal Ministry of Education and Research Germany under the projects DECODE (01IW21001) and COPPER (01IW24009) and partially under the EU project ExtremeXP (GA Nr 101093164).

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *ECCV*, 2018. 2

[2] Francesco Barbato, Elena Camuffo, Simone Milani, and Pietro Zanuttigh. Continual road-scene semantic segmentation via feature-aligned symmetric multi-modal network. *arXiv preprint arXiv:2308.04702*, 2023. 3

[3] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulò, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *CVPR*, 2020. 2

[4] Sungmin Cha, YoungJoon Yoo, Taesup Moon, et al. Ssul: Semantic segmentation with unknown label for exemplar-based class-incremental learning. *Advances in Neural Information Processing Systems (NeurIOS)*, 2021. 2

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv:1802.02611*, 2018. 6

[6] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *ECCV*, 2020. 3

[7] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. Indoor semantic segmentation using depth information: 1st international conference on learning representations, iclr 2013. In *1st International Conference on Learning Representations, ICLR 2013*, 2013. 3

[8] Liuyuan Deng, Ming Yang, Tianyi Li, Yuesheng He, and Chunxiang Wang. Rfbnet: deep multimodal networks with residual fusion blocks for rgb-d semantic segmentation. *arXiv preprint arXiv:1907.00135*, 2019. 3

[9] Yao Deng and Xiang Xiang. Replaying styles for continual semantic segmentation across domains. In *Pattern Recognition: Asian Conference*, 2023. 2

[10] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *CVPR*, 2021. 2

[11] Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017. 2

[12] Gianni Franchi, Marwane Hariat, Xuanlong Yu, Nacim Belkhir, Antoine Manzanera, and David Filliat. Infraparis: A multi-modal and multi-task autonomous driving dataset. In *WACV*, 2024. 3, 5, 7, 8, 11, 12

[13] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, 2018. 2

[14] Prachi Garg, Rohit Saluja, Vineeth N Balasubramanian, Chetan Arora, Anbumani Subramanian, and CV Jawahar. Multi-domain incremental learning for semantic segmentation. In *WACV*, 2022. 2

[15] Alexander Gepperth and Cem Karaoguz. A bio-inspired incremental learning architecture for applied perceptual problems. *Cognitive Computation*, 2016. 2

[16] Dipam Goswami, René Schuster, Joost van de Weijer, and Didier Stricker. Attribution-aware weight transfer: A warm-start initialization for class-incremental semantic segmentation. In *WACV*, 2023. 2

[17] Tyler L Hayes, Nathan D Cahill, and Christopher Kanan. Memory efficient experience replay for streaming learning. In *International Conference on Robotics and Automation (ICRA)*, 2019. 2

[18] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *ACCV*, 2017. 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[20] Xu He and Herbert Jaeger. Overcoming catastrophic interference by conceptors. *arXiv preprint arXiv:1707.04853*, 2017. 2

[21] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *CVPR*, 2019. 2

[22] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *ICIP*, 2019. 3

[23] Herbert Jaeger. Using conceptors to manage neural long-term memories for temporal patterns. *Journal of Machine Learning Research*, 2017. 2

[24] Heechul Jung, Jeongwoo Ju, Minju Jung, and Junmo Kim. Less-forgetting learning in deep neural networks. *arXiv preprint arXiv:1607.00122*, 2016. 2

[25] Prakhar Kaushik, Alex Gain, Adam Kortylewski, and Alan Yuille. Understanding catastrophic forgetting and remembering in continual learning with optimal relevance mapping. *arXiv preprint arXiv:2102.11343*, 2021. 2, 4, 6, 7, 8, 11, 12

[26] Ronald Kemker and Christopher Kanan. Fearnet: Brain-inspired model for incremental learning. *arXiv preprint arXiv:1711.10563*, 2017. 2

[27] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 2017. 2, 6, 7, 11, 12

[28] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *CVPR*, 2018. 2

[29] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*. Elsevier, 1989. 1

[30] Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in Psychology*, 2013. 1

[31] Umberto Michieli and Pietro Zanuttigh. Knowledge distillation for incremental learning in semantic segmentation. *Computer Vision and Image Understanding*, 2021. 2, 6, 7, 11, 12

[32] Hazem Rashed, Ahmad El Sallab, Senthil Yogamani, and Mohamed ElHelw. Motion and depth augmented semantic segmentation for autonomous navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 3

[33] Nikhil Reddy, Mahsa Baktashmotlagh, and Chetan Arora. Towards domain-aware knowledge distillation for continual model generalization. In *WACV*, 2024. 2

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015. 6

[35] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 2

[36] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *NeurIPS*, 2017. 2

[37] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 2019. 3

[38] Yuxiang Sun, Weixun Zuo, Peng Yun, Hengli Wang, and Ming Liu. Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion. *IEEE Transactions on Automation Science and Engineering*, 2021. 3

[39] Johan Vertens, Jannik Zürn, and Wolfram Burgard. Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. *arXiv preprint arXiv:2003.04645*, 2020. 5, 6, 7, 8, 12

[40] Yikai Wang, Fuchun Sun, Ming Lu, and Anbang Yao. Learning deep multimodal feature representation with asymmetric multi-layer fusion. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 3

[41] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Growing a brain: Fine-tuning by increasing model capacity. In *CVPR*, 2017. 2

[42] Chenshen Wu, Luis Herranz, Xialei Liu, Joost Van De Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. *NeurIPS*, 2018. 2

[43] Shi Yi, Junjie Li, Xi Liu, and Xuesong Yuan. Ccaffmnet: Dual-spectral semantic segmentation network with channel-coordinate attention feature fusion module. *Neurocomputing*, 2022. 3

[44] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *ICLR*, 2018. 2

[45] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *ICML*, 2017. 2

[46] Chang-Bin Zhang, Jia-Wen Xiao, Xialei Liu, Ying-Cong Chen, and Ming-Ming Cheng. Representation compensation networks for continual semantic segmentation. In *CVPR*, 2022. 2

[47] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 3

[48] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In *CVPR*, 2021. 3

[49] Wujie Zhou, Jinfu Liu, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Gmnet: Graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation. *IEEE Transactions on Image Processing*, 2021. 3

Supplementary Material

A. Overview

In this supplementary material to our paper *Modality-Incremental Learning with Disjoint Relevance Mapping Networks for Image-based Semantic Segmentation*, we show the impact of forgetting on previously learned modalities, test the robustness of Disjoint Relevance Mapping Networks (DRMNs) against variation of the pruning parameter μ, and list the exact utilization of network connections for the experiment on InfraParis [12].

B. Task-wise Evaluation

To quantify the amount of forgetting due to the incremental learning of different modalities, Tab. 5 provides the mIoU for each modality through the learning sequence. *I.e.*, each known modality is evaluated after each task. This way, the mutual negative influence of the modalities can be measured. With regularization-based approaches such as EWC [27] and ILT [31], the model learns optimally during the initial step as expected. However, when learning other modalities incrementally, EWC prevents overwriting important parameters from the previous modalities, hindering its learning on the new modality. On the other hand, ILT which uses distillation, exhibits better performance on the initial task compared to EWC. However, the performance on new modalities is significantly worse due to the diverse nature of the modalities. In RMN [25] and the proposed DRMN, even for the initial modality the results are slightly lower compared to the single-task models. This is due to the use of relevance maps, which preserve network capacity for future tasks by not utilizing the entire network capacity at each step. This approach effectively preserves information and completely mitigates catastrophic forgetting, ensuring that performance on previously learned modalities remains consistent over the sequence of tasks. Additionally, with DRMN, isolating parameters and learning task-specific weights enhances the learning of new modalities, as evident in improved performance on both Gray and RGB tasks.

C. Relevance Map Pruning

To recall, the hyperparameter μ defines the threshold at which network weights (connections) are considered relevant. Any connection below this threshold will be pruned after every epoch above 50. The values in the relevance map of the pruned connections will be permanently set to zero for this task, removing the influence of that connection entirely. The unused connections might be used in a later task, though. The network's weights for relevant connections will be frozen. The default value for μ is 0.6. In Tab. 6, we show the results for a threshold of 0.5 and 0.7. The variation of μ in both directions indicates a high robustness of DRMN in this regard. Same holds for the original RMN. Interestingly, we point out that varying the pruning parameter has no significant impact on the sparsity (utilization) of neural connections.

D. Task Utilization on InfraParis

One of our claims in the main paper is that despite the strict separation of task-specific connections, the network's capacity is not exceeded faster than with regular RMNs. To back this claim further, we have also computed the network utlization for the four tasks of InfraParis [12]. The result is shown in Tab. 7. For a description of ORMN and PRMN, we refer to Sec. 5.5 of the main paper. It is striking that even with just about 6 % of the network's overall connections, the final task can be learned even better than with RMN, which uses about half of all weights. Another remarkable observation is that each task approximately consumes half of the remaining connections in DRMN.

Table 5. Results on Freiburg Thermal [39] for the original RMN [25] and our proposed DRMN after each step of training the sequence (***IR*** $\rightarrow$ ***Gray*** $\rightarrow$ ***RGB***).

Method	M_0 (IR)	M_1 (Gray)		M_2 (RGB)		
	IR	IR	Gray	IR	Gray	RGB
Fine Tuning	59.56	07.44	74.10	07.12	73.17	75.24
EWC [27]	59.75	06.89	58.04	10.56	59.47	61.69
ILT [31]	59.56	20.39	21.08	20.68	22.57	23.53
RMN [25]	55.30	55.18	68.85	55.10	68.90	69.46
DRMN (Ours)	55.30	55.16	70.61	54.97	70.56	71.19

Table 6. Results and task-wise network utilization on Freiburg Thermal [39] for the original RMN [25] and our proposed DRMN with varying pruning parameters.

Method	**Prune**	**Results (mIoU)**				**Task Utilization (%)**			**Overall (%)**	
	μ	IR	Gray	RGB	**Avg**	IR	Gray	RGB	Shared Weights	Network Utilization
RMN [25]	0.5	55.02	69.06	68.96	64.35	49.91	49.85	49.83	49.80	87.39
	0.6	55.10	68.90	69.46	64.49	47.79	47.78	47.80	46.68	85.77
	0.7	54.25	68.33	68.66	63.75	49.10	49.03	49.08	48.62	86.78
DRMN (Ours)	0.5	55.23	70.64	71.21	65.69	49.91	24.94	12.50	0.00	87.35
	0.6	54.97	70.56	71.19	65.57	47.79	24.95	13.03	0.00	85.78
	0.7	54.34	70.69	71.05	65.36	49.10	24.93	12.73	0.00	86.76

Table 7. Network utilization on InfraParis [12] for the original RMN [25] and our proposed DRMN on the task sequence (***IR*** $\rightarrow$ ***RGB*** $\rightarrow$ ***Depth*** $\rightarrow$ ***Gray***).

Method	**Task Utilization (%)**				**Overall (%)**	
	IR	RGB	Depth	Gray	Shared Weights	Network Utilization
RMN [25]	49.52	49.54	49.49	49.54	68.02	93.50
ORMN	49.52	49.46	49.40	49.41	67.91	93.47
PRMN	49.52	27.16	15.84	10.16	5.97	93.48
DRMN (Ours)	49.52	24.98	12.61	6.37	0.00	93.49