OGP-Net: Optical Guidance Meets Pixel-Level Contrastive Distillation for Robust Multi-Modal and Missing Modality Segmentation

Aniruddh Sikdar[1], **Jayant Teotia**[1], **Suresh Sundaram**[2]

[1]Robert Bosch Centre for Cyber-Physical Systems, Indian Institute of Science, Bangalore
[2]Department of Aerospace Engineering, Indian Institute of Science, Bangalore
aniruddhss@iisc.ac.in, jayantteotia@iisc.ac.in, vssuresh@iisc.ac.in

Abstract

Enhancing the performance of semantic segmentation models with multi-spectral images (RGB-IR) is crucial, particularly for low-light and adverse environments. While multi-modal fusion techniques aim to learn cross-modality features for generating fused images or engage in knowledge distillation, they often treat multi-modal and missing modality scenarios as separate challenges, which is not an optimal approach. To address this, a novel multi-modal fusion approach called Optically-Guided Pixel-level contrastive learning Network (OGP-Net) is proposed, which uses Distillation with Multi-View Contrastive (DMC) and Distillation for Uni-modal Retention (DUR) to maintain the correlation between modality-shared and modality-specific features. DMC aligns the uni-modal features by projecting the semantic information across modalities into a unified latent space, ensuring that the feature maps retain multi-modal representations. Pixel-level multi-view contrastive learning is introduced to enable modality-invariant representation learning. To retain modality-specific information, DUR is proposed, which distills detailed textures from RGB images into the optical branch of OGP-Net. Additionally, the Gated Spectral Unit (GSU) is integrated into the framework to eliminate the need for manual tuning and avoid forced feature alignment. Comprehensive experiments show that OGP-Net outperforms state-of-the-art models in multi-modal and missing modality scenarios across three public benchmarking datasets. It achieves quicker convergence and learns efficiently from limited training samples.

Introduction

Recent advancements in scene parsing have achieved impressive segmentation results (Udupa et al. 2024), but these improvements are largely specific to RGB data (Sikdar et al. 2023). Enhancing models' generalization under challenging conditions like low light, or overexposure, where information contained in RGB images depreciates, remains crucial. Infrared (IR) cameras are increasingly popular in low-visibility conditions for their unique spectral information and ability to penetrate dust and smoke (Gade and Moeslund 2014). However, deep-learning models often struggle with IR images alone due to their lower semantic content compared to optical images. With the rise of affordable IR sensors, deep multimodal fusion (Valada, Mohan, and Burgard 2020) has gained traction, combining RGB and IR modalities (Ha et al. 2017; Kütük and Algan 2022) to improve semantic segmentation and outperform unimodal approaches. The primary focus is capturing shared information across modalities while preserving each modality's unique semantic knowledge.

Multimodal systems observe the same object through different imaging systems, where low frequencies capture shared information and high frequencies highlight each modality's distinct features (Zhao et al. 2023a). For example, RGB images reveal texture details, while IR images highlight thermal radiation. Fusion techniques for combining RGB and IR modalities fall into two main categories: (1) *Multi-Modal Image Fusion (MMIF)* and (2) *Multi-Modal Feature-Level Fusion (MMFF)*. MMIF techniques (Liu et al. 2022) focus on generating fused images by modeling cross-modality features from different sensors. A common pipeline using auto-encoders (Liang et al. 2022) to fuse RGB and IR images often neglects modality-specific high-frequency information (Sener and Koltun 2018; Liang et al. 2021). CDDFuse (Zhao et al. 2023a) addresses this by leveraging correlations between low and high frequencies within the image space to constrain the solution space. Despite the superior image quality metrics of MMIF models, these correlations are not fully explored during model training, leading to performance degradation.

To deal with the MMIF model's challenges in handling missing modality scenarios, MMFF techniques involving (i) *feature-level fusion*, and (ii) *knowledge distillation* are explored to align different sensor modality distributions in the feature space. (i) *Feature-level fusion* techniques have been proposed, as shown in Fig. 1(a), using separate encoders for each modality (Wei, Luo, and Luo 2023), followed by attention-based modules (Yu et al. 2020) leading to a significant increase in parameters. However, they do not exhibit superior performance compared to the parameter-free feature exchange techniques (Wang et al. 2022). Channel Exchange Network (Wang et al. 2022) facilitates message passing in both encoders by dynamically exchanging channels to enable the integration of information. (ii) *Knowledge distillation* techniques are being developed to transfer pixel-level semantic information from RGB data to other modalities. However, for modalities with substantial domain gaps, directly aligning modality-specific features can lead to nega-

tive transfer due to the forced feature alignment (Kang et al. 2022). Most previous work treats the challenges of multi-modal and missing modality issues as distinct and separate problems.

To address this, we propose a novel **O**ptically-**G**uided **P**ixel-level contrastive learning **N**etwork (OGP-Net), which distills knowledge from a pre-trained RGB model for multi-modal segmentation. The correlation between modality-shared and modality-specific features is managed in the feature space using the proposed Distillation with Multi-View Contrastive (DMC) and Distillation for Unimodal Retention (DUR). In OGP-Net, encoders use shared convolutional filters and distinct batch normalization layers to project rich semantic knowledge from both modalities into a unified latent space. DMC is a simple and self-adaptive mechanism that enhances the correlation between modality-shared features, generating multi-view feature maps. Pixel-level contrastive learning is then applied to these maps, aligning RGB-IR features by contrasting similar and dissimilar samples, as shown in Fig. 1 (b). This approach improves the model's ability to effectively integrate and interpret information from diverse sources. DUR manages the correlation of high-frequency features by distilling knowledge to the optical branch (RGB branch) of OGP-Net. The Gated Spectral Unit (GSU) is introduced to integrate information from RGB, IR, and fused predictions, enabling efficient knowledge distillation. This leads to superior performance in multi-modal and missing-modality scenarios while maintaining similar model complexity to the baseline model. The main contributions of this paper can be summarized as follows:

- We introduce OGP-Net, a novel multi-spectral semantic segmentation network. This network utilizes pixel-level optically-guided knowledge distillation to transfer multi-level semantic features from RGB images to the OGP-Net using Distillation with Multi-View Contrastive (DMC) and Distillation for Uni-modal Retention (DUR).
- The framework employs a novel Distillation with Multi-View Contrastive (DMC) strategy which balances intra- and inter-modal semantic propagation. This approach helps increase the constraints on the correlation of modality-shared features using pixel-level multi-view contrastive learning.
- Experimental evaluations on three public datasets show that OGP-Net consistently outperforms state-of-the-art models, including those for multimodal fusion and missing modality, particularly in challenging conditions like low light and adverse weather.
- A feature reuse strategy is employed to maintain baseline computational complexity while boosting performance for infrared images (for missing-modality scenarios). Additionally, OGP-Net is designed to train effectively with limited co-registered RGB-IR datasets and achieves fast convergence, making it suitable for resource-constrained devices.

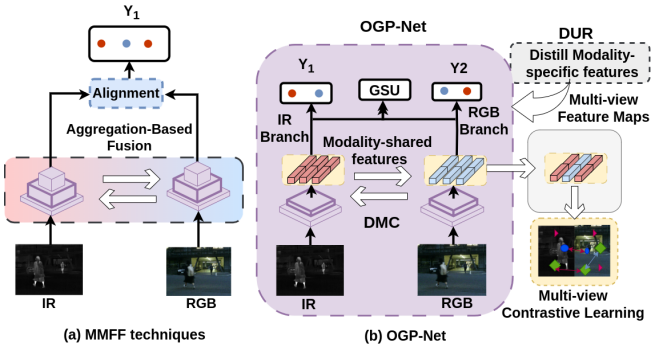

Figure 1: Comparison between existing fusion techniques (MMFF) and the proposed OGP-Net. Distillation with Multi-View Contrastive (DMC) and Distillation for Unimodal Retention (DUR) are proposed to leverage modality-shared and specific information.

Related Works

Deep multi-modal fusion aims to capture fine-grained details from multiple imaging sources to address uni-modal defects (He 2024). MMIF methods aim to generate a fused image that retains characteristics from both modalities. CDDFuse (Zhao et al. 2023a) is one such approach, which uses Restormer blocks to extract low-level features from each modality. MMFF techniques can be broadly categorized into aggregation-based and alignment-based fusion techniques. Aggregation-based techniques use operators to merge multi-modal images into a single network, as shown in Fig. 1(a), often with attention-based modules. For instance, the Cross-Modality Transformer (Qingyun, Dapeng, and Zhaokui 2021) captures dependencies across data and integrates global context. Recent methods (Zhang et al. 2023) further enhance this by using separate subnetworks for each modality. MMA-Net (Wei, Luo, and Luo 2023) introduces a framework where the teacher network transfers comprehensive multimodal information to the deployment network, improving its ability to handle weaker modality combinations. Due to the scarcity of large-scale, co-registered RGB-IR datasets, data-intensive deep learning models face overfitting issues (He 2024). Alignment methods, like the Channel Exchanging Network (C.E.N) (Wang et al. 2022), propose dynamically swapping channels between sub-networks for information fusion, which can reduce the discriminability of the fused features and poor performance in scenarios with missing modalities.

Knowledge Distillation (KD) focuses on transferring knowledge from a complex neural network to a smaller model (Hinton, Vinyals, and Dean 2015), aiding deployment on resource-limited devices. Pixel-level distillation techniques like Cross-Image Relational Knowledge Distillation (CIRKD) (Yang et al. 2022) transfer structured pixel and region relationships using a memory bank but requires increased training complexity. Adaptive Perspective Distillation (APD) uses two projection heads to perform pixel-level distillation, minimizing KL divergence between the teacher's and student's projected feature embeddings (Tian

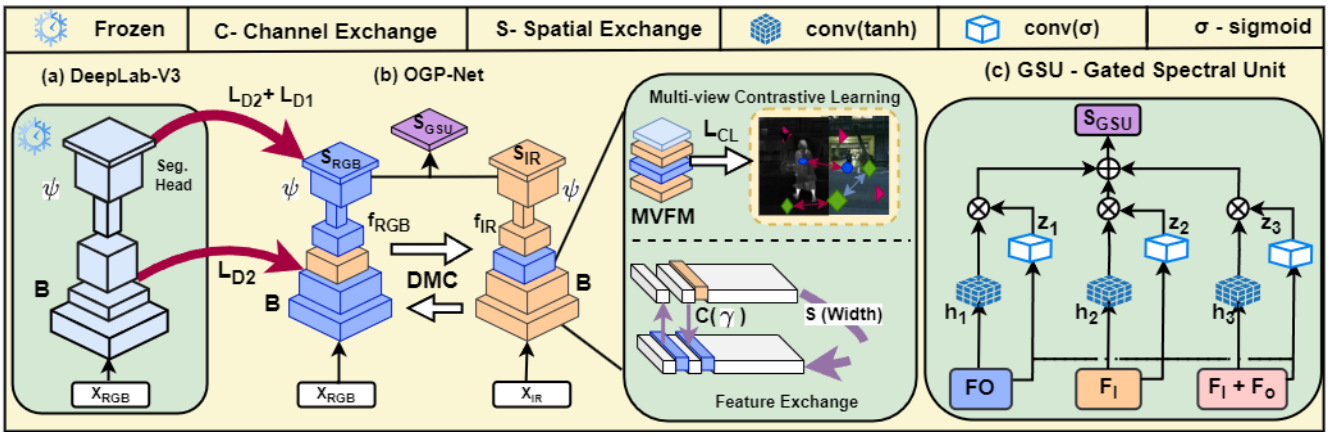

Figure 2: Network architecture of OGP-Net: The model receives inputs $\{X_{RGB}, X_{IR}\}$. The DMC strategy is implemented to improve modality-shared information by utilizing feature exchange and simultaneous distillation from (a) frozen baseline DeepLabV3+ model. Multi-View Feature Maps (MVFM) are generated and subsequently trained using supervised, pixel-level contrastive learning. (c) Additionally, the Gated-Spectral Unit (GSU) is incorporated to eliminate the need for manual adjustments. (Dotted lines indicate that they are not physically connected.)

et al. 2022). Vanilla KD with KL divergence or CWD losses from a pre-trained optical (RGB) model can disrupt feature alignment, especially with high discrepancies between teacher and student models. Cross-modality distillation is crucial in deep learning for transferring knowledge from one modality to enhance those with limited information, such as depth maps and high-quality sketches. However, extending these methods to modalities with significant feature discrepancies, such as RGB-IR (Do et al. 2024), remains an emerging area. For multi-sensor applications, novel techniques are needed to distill knowledge effectively while keeping the model complexity low.

Contrastive Learning is an unsupervised representation learning method that helps in learning distinct feature representations by distinguishing between similar and dissimilar pairs (Chen et al. 2020). Multi-view contrastive learning has been introduced for classification tasks, exploring the use of images from different sensors at the same location and time as positive examples to maximize mutual information, thus eliminating the need for manual augmentation tuning (Berg et al. 2024). Pixel-level contrastive learning techniques has been mainly used in self-supervised, domain adaptation (Chen et al. 2023) and class-incremental settings (Zhao, Yuan, and Shi 2023). Pixel-wise contrastive distillation (Huang and Guo 2023) introduces a self-supervised framework that aligns positive pixel pairs and repulses negative pairs to match the distributions of corresponding pixels in student and teacher feature maps. These techniques require a memory buffer, leading to significant memory overhead for augmented samples and feature maps. Multi-view contrastive learning for pixel-level scenarios remains an active research area (Jain, Wilson, and Gulshan 2022). Traditional methods using separate projection heads for each modality are inefficient, as they lose spatial information and struggle with feature matching when discrepancies are high. In contrast, our approach explores core concepts of supervised multi-view pixel-level contrastive learning within the latent space, utilizing multi-view feature maps to effectively mitigate semantic drift.

Methodology

Problem Formulation

Multi-modal segmentation models aim to bridge the cross-modal gaps between RGB and IR modalities while ensuring robustness in both multi-modal and missing modality scenarios. Let $\{X_{RGB}, X_{IR}\}$ denote the co-registered RGB-IR image pairs from dataset *D*, along with their corresponding pixel-wise labels *Y*. Data from both modalities are passed as input to the semantic segmentation model F_θ:$\{X_{RGB}, X_{IR}\} \rightarrow \mathbf{S}$, where **S** is the segmentation mask and θ represents the learnable parameters. During inference, for multi-modal settings, $\{X_{RGB}, X_{IR}\}$ images are passed to the model. Only the IR data $\{X_{IR}\}$ is passed to the model for missing modality scenarios. OGP-Net takes a unified approach by addressing both problems concurrently rather than treating them separately.

Overview

Segmentation model *M* can be decomposed into ψ o *B*, where *B* is the backbone network, and ψ represents the decoder and segmentation head. The backbone layers capture low-level, modality-shared features by preserving local structures (Zeiler and Fergus 2014; Kang et al. 2022). OGP-Net inputs $\{X_{RGB}, X_{IR}\}$ into the backbone *B*, i.e, *B*: $\{X_{RGB}, X_{IR}\} \rightarrow \{f_{RGB}, f_{IR}\}$. *B* is the proposed Distillation with Multi-view Contrastive (DMC) block, as shown in 2. The encoder blocks (backbone) consist of five stages, $\{f^i | i = 1, 2, 3, 4, 5\}$, where the i^{th} encoding stage is denoted as $f^i(.)$, and maps the features F_{i-1} to F_i. As the model progresses through its layers, the features tend to become more specific to particular modalities. Hence, these

later layers are preserved individually to retain the modality-specific semantic information. The output of the backbones is designed to be fed into separate decoders for each modality, generating both individual segmentation heads and a unified fused segmentation head output, jointly denoted as: ψ: $\{f_{RGB}, f_{IR}\} \rightarrow \{S_{RGB}, S_{IR}, S_{GSU}\}$.

OGP-Net: Optically-Guided Pixel-level contrastive learning Network

Training scheme: The training process consists of two key steps: (1) pre-training the baseline DeepLabV3+ segmentation model on RGB images, and (2) training the OGP-Net model using RGB-IR images while concurrently distilling knowledge from the pre-trained model into the optical (RGB) branch of OGP-Net.

Training step 1: During this step, the baseline segmentation model is trained with the RGB images with their corresponding labels $\{X_{RGB}, Y\}$ to generate a well-structured pixel embedding space. An auxiliary head (h_{aux}) is used to project the features from the 4^{th} layer of the encoder, and train with downsampled labels $\{y^D\}$. This is to train a more robust encoder, by emphasizing the learning of the low-frequency, modality-shared features within the model. Segmentation loss $L_{seg}(p, y)$ is used to train both the model and the auxiliary head (h_{aux}). It consists of the summation of cross-entropy and dice loss, and is given by,

$$L_{seg}(p, y) = -\sum_i y_i log(p_i) + 1 - \frac{2\sum_i p_i y_i}{\sum_i y_i + \sum_i p_i} \quad (1)$$

where *y* and *p* denote the ground truth labels and the pixel-wise predictions, respectively. Once trained, the baseline model is kept frozen.

Training step 2: During the second training step, the OGP-Net (student) model undergoes training with pixel-level distillation from the pre-trained RGB model (teacher) using the proposed framework, as shown in Fig. 2 (b). This process primarily involves Distillation with Multi-View Contrastive (DMC) and Distillation for Unimodal Retention (DUR). DMC balances intra- and inter-modal learning by projecting semantic information across modalities into a unified latent space, enabling modality-invariant representation learning. In contrast, DUR focuses on preserving modality-specific representations within the latent space.

Distillation with Multi-view Contrastive (DMC) To capture common semantics from both the modalities, the proposed DMC integrates (i) *feature exchange*, (ii) *knowledge distillation*, and (iii) *multi-view contrastive learning* simultaneously, serving as a self-adaptive fusion method in the dual branches of the OGP-Net encoder. Convolutional layers capture modality-shared features across different sensors, while batch normalization layers are crucial for preserving modality-specific characteristics (Zheng et al. 2021; Chang et al. 2019; Wang et al. 2020). Hence, the dual branches of the encoder in OGP-Net share convolutional filters across both modalities while maintaining distinct batch normalization layers for each modality.

(i) Feature exchange: Batch normalization performs normalization on feature maps, followed by affine transformation using γ and β parameters. The correlation between the feature map f_i and its corresponding scaling factor γ_i on output predictions is well studied, revealing redundancy in feature maps when $\gamma \rightarrow 0$ (Liu et al. 2017; Ye et al. 2018). The feature exchange mechanism (Wang et al. 2022; De Vries et al. 2017; Zhang et al. 2018) has been employed for cross-modal message passing by modulating batch normalization. This mechanism has been used to dynamically exchange channels between the dual branches of the encoder, embedded with the same mapping of RGB-IR modalities, due to shared convolutional filters. This approach helps align the unimodal features, ensuring that the feature maps retain multi-modal representations from both modalities. Channel exchange *C* is performed across all stages of the encoders within OGP-Net. Channels are interchanged along the batch axis in the encoder when the scaling factor γ is close to 0. Spatial exchange *S* is also introduced to exchange features in the dual branches, specifically in the spatial dimension. An exchange mask denoted as $\mathrm{M} \in R^{n,c,h,w}$ is generated, wherein values of 0 and 1 correspond to elements designated for non-exchange and exchange, as given below,

$$M(n, c, h, w) = \begin{cases} 0 & \text{if} \quad w\%2 = 0 \\ 1 & \text{if} \quad otherwise \end{cases} \quad (2)$$

where *n*, *c*, *h*, and *w* represent the batch size, number of channels, and height width respectively. Features along the width dimension are exchanged in the dual branch only for the last two stages of the encoders. Both channel and spatial exchanges are avoided in deeper layers to prevent interference with modality-specific propagation, as shown in Fig. 2 (b). These feature exchanges facilitate a robust knowledge transfer between the dual branches, allowing the feature maps to retain multi-modal representations from both modalities.

(ii) Pixel-lexel knowledge distillation: Given the significant discrepancy between RGB and IR modalities, strict constraints are enforced to align the pixel-wise feature maps between the pre-trained model and OGP-Net. The multi-level semantic information is distilled from the last two layers of the encoder of the pre-trained model (F^{PO}) to the last two layers of the shared encoders of the optical branch of OGP-Net, using the mean square error loss, given by,

$$L_{D2}(F, F^{PO}) = \sum_{i \in \{4,5\}} \|F_i - F_i^{PO}\|_2 + \|F_d - F_d^{PO}\|_2 \quad (3)$$

which measures the difference between the features of the last two layers of the encoders (F_i) and the decoder output (F_d). This distillation facilitates the transfer of knowledge from the well-structured pixel embedding space of the pre-trained model to OGP-Net.

(iii) Multi-view contrastive learning: OGP-Net's shared encoders project semantic knowledge from various perspectives into a unified latent space. Feature exchange and distillation align cross-modal features, transforming unimodal features into rich multi-modal representations, known as multi-view feature maps. We propose training the last four encoder layers with these multi-view feature maps using pixel-level contrastive learning L_{CL} (Gutmann and Hyvärinen 2010) to enhance intra-domain mining, as shown

in Fig. 2 (b). This approach reduces the need for additional augmentations or explicit views from different sensors, as the feature maps encapsulate modality-shared information from both sensors. Features are mapped into an embedding space using a single projection head (a non-linear MLP h_{pixel}), bringing pixel embeddings of the same category closer together while pushing those from different categories farther apart. Pixel-wise labels are used to categorize pixels of the same class as positive samples and those of different classes as negative samples. The pixel-wise contrastive loss is formulated as,

$$L_{CL} = -\sum_{C(i)=C(j)} log\frac{r(e_i, e_j)}{\sum_{k=1}^{Np} r(e_i, e_j)} \quad (4)$$

where, e_i represents the i^{th} feature map obtained from the projection head, *Np* stands for the total number of pixels, *r* denotes the similarity measure. Similarity is calculated using the exponential similarity function: $r(e_i, e_j) = exp(s(e_i, e_j) / \tau)$, where s represents the cosine similarity, and τ is the temperature parameter. Various sampling strategies can be employed, but empirically, the semi-hard example sampling strategy yields the best performance. In this strategy, for each anchor, the top 10% of the closest negative samples and the most distant positive samples are selected for each anchor (Chen et al. 2023).

Sparsity regularization is used on batch normalization scaling factors to automatically eliminate redundant channels, resulting in more compact models (Liu et al. 2017). This method is crucial for models with feature exchange mechanisms and has been shown to significantly boost performance (Shao et al. 2020). However, in training OGP-Net, pixel-level knowledge distillation leads to frequent changes in feature-level statistics, causing constant fluctuations in batch normalization parameters, especially γ. As a result, the sparsity constraints become overly restrictive and are not applied during training (more details in supplementary material).

Distillation for Unimodal Retention (DUR) The multi-view features from the backbones are fed into the decoder function to produce output predictions for each modality, as well as a fused segmentation head output, i.e., ψ: $\{f_{RGB}, f_{IR}\} \rightarrow \{S_{RGB}, S_{IR}, S_{GSU}\}$. To preserve high-frequency, modality-specific information, we propose using the DUR strategy in the decoder and segmentation heads, which integrates knowledge distillation and feature fusion.

To align the output predictions of the optical branch of OGP-Net (S_{RGB}) with the output of the pre-trained model, we introduce the DIST loss (Huang et al. 2022) into the framework. It distills the inter-class knowledge and captures the intra-relation of semantic similarities from the multi-class pixel-wise predictions obtained from the pre-trained model, denoted as p^{PO}. The distillation loss L_{D1} is the sum of L_{inter} and L_{intra} losses, defined as:

$$L_{inter}(S_{RGB}, p^{PO}) = \frac{1}{B}\sum_i d_p(S_{RGB(i,:)}, p^{PO}_{(i,:)}) \quad (5)$$

$$L_{intra}(S_{RGB}, p^{PO}) = \frac{1}{C}\sum_j d_p(S_{RGB(:,j)}, p^{PO}_{(:,j)}) \quad (6)$$

where *B* and *C* represent the rows and columns of the prediction feature maps. The similarity metric $d_p(.,.)$ is defined as $d_p(.,.) = 1 - \rho(.,.)$, where $\rho(.,.)$ is the Pearson correlation coefficient between two sets of output logits.

Gated Spectral Unit (GSU) (Arevalo et al. 2017) is introduced to merge similar visual concepts from both imaging modalities and produce the fused output prediction S_{GSU}. It uses gating mechanisms to automatically determine the influence of different units' activations, eliminating the need for manual adjustments. Fig. 2 (c) depicts the structure of a GSU. The output of the RGB and IR branches and their summation are fed to the GSU block. We aggregate the feature maps from both modalities using a summation operation to enhance high-frequency features. These features maps are first passed through convolution layers and through the layers as given below,

$$h_1 = tanh(W_1 * F_I), \quad (7)$$

$$h_2 = tanh(W_2 * F_O), \quad (8)$$

$$h_3 = tanh(W_3 * (F_I + F_O)) \quad (9)$$

where, $\{W_1, W_2, W_3\}$ represents the convolution weights. For each branch, gate neuron Z is computed, given by,

$$Z_1 = \sigma(W_{1'} \otimes [F_I, F_O, F_I + F_O]) \quad (10)$$

where $[\cdot, \cdot]$ denotes the concatenation operator and σ denotes sigmoid operation. Z_2 and Z_3 are also calculated in a similar manner to Z_1. The final output predictions of the fusion block S_{GSU} is given by,

$$S_{GSU} = Z_1 \otimes h_1 + Z_2 \otimes h_2 + Z_3 \otimes h_3 \quad (11)$$

where $(\otimes)$ represents the multiplication operation. OGP-Net model has three outputs as shown in the figure, two for each modality, i.e., for RGB and IR, and the third output from the GSU block, given by S_{GSU}. GSU helps in preserving the high-frequency, modality-specific information within the IR branch of the model during distillation, by preventing forced feature alignment.

Optimization As the data propogates throughout the model, various loss functions are applied at intermediate layers to optimize the learning process. The joint loss function for training OGP-Net is given by,

$$L_{ST2} = L_{seg}(S_{(GSU,IR,RGB)}, y) + L_{seg}(p_{aux}, y^D) + L_{D1}(S_{RGB}, p^{PO}) + L_{D2}(F, F^{PO}) + L_{CL} \quad (12)$$

where y and y^D represent the ground truth labels and the downsampled labels, respectively. p_{aux} denotes the features projected using the auxiliary head (h_{aux}). L_{seg} is used to optimize the model's output from the fused segmentation predictions S_{GSU} generated using GSU, as well as the RGB and IR output predictions S_{RGB} and S_{IR}. During inference, for multi-modal settings, the final predictions from the GSU block are used. In the missing modality scenarios, there are two settings: (1) the final predictions from the GSU block are used, and (2) only the IR branch of OGP-Net is employed (S_{IR}), referred to as OGP-Net-IR. OGP-Net-IR has the same model parameters as that of DeepLabV3+.

Multi-Spectral Settings	Method	MSRS (mIoU)	MVSS (mIoU)	Params (M)
Baseline	RGB (Chen et al. 2018)	64.33	48.89	11.68M
	IR (Chen et al. 2018)	61.96	42.82	11.68M
Multi-Modal Settings (RGB-IR)	TarDAL (Liu et al. 2022)	65.80	21.21	11.71M
	CDDFuse (Zhao et al. 2023a)	66.71	48.41	13.47M
	DDFM (Zhao et al. 2023b)	61.62	-	564.494M
	C.E.N (Wang et al. 2022)	61.04	51.33	99.13M
	MMANet (Wei, Luo, and Luo 2023)	66.24	49.31	71.70M
	OGP-Net (Ours)	**69.97**	**52.90**	14.24M
Missing Modality (IR only)	DisOptNet (Kang et al. 2022)	63.84	43.22	11.68M
	MMANet (Wei, Luo, and Luo 2023)	61.34	46.83	50.69M
	C.E.N (Wang et al. 2022)	62.93	40.33	99.13M
	SKD-Net (Sikdar, Teotia, and Sundaram 2024)	64.67	45.53	11.68M
	OGP-Net (Ours)	**66.77**	**47.18**	14.24M

Table 1: Performance comparison of mean Intersection over Union (mIoU) of proposed OGP-Net with other state-of-the-art models on MSRS and MVSS datasets. OGP-Net outperforms other models in challenging low-light scenarios with significantly fewer parameters.

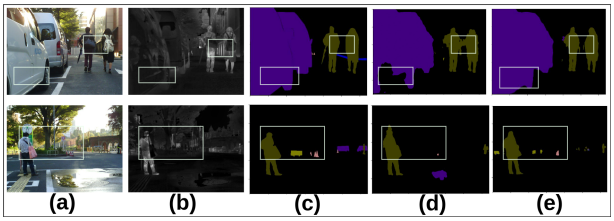

Figure 3: Comparison of output predictions for the MSRS dataset for missing modality scenario, with (a) RGB, (b) IR inputs, (c) labels, and predictions from (d) DisOptNet and (e) OGP-Net. (See supplementary material for additional results.)

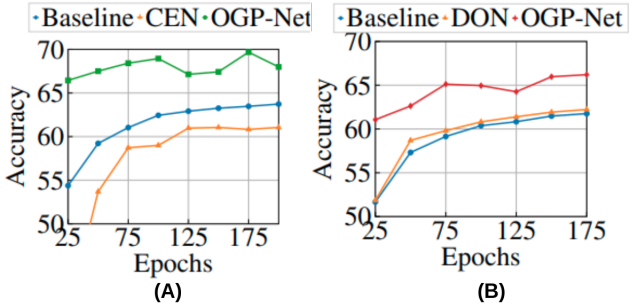

Figure 4: Performance curves of OGP-Net and other models for (A) multi-modal and (B) missing-modality scenarios on MSRS dataset. OGP-Net demonstrates faster convergence compared to other models.

Experimental Results

Datasets and implementation details Three public datasets are used for benchmarking several downstream multi-modal segmentation tasks: MSRS (Tang et al. 2022), MVSS (Ji et al. 2023), and FMB (Liu et al. 2023). These datasets contain images of urban scenes, with a wide variety of environmental conditions. Intersection over Union (% mIoU) is used as a metric to quantitatively measure the segmentation performance. DeepLabV3+ (Chen et al. 2018) with the EfficientNet-B3 (Tan and Le 2019) backbone serves as the baseline segmentation model. All models are trained with a batch size of 8 for 200 epochs, and data augmentation includes horizontal flips with a 50% probability. All experiments are conducted on NVIDIA Quadro RTX 5000 GPUs, and which is pre-trained on ImageNet. The models are trained using an SDG optimizer, with an initial rate of 5 $\times 10^{-3}$.

Comparison with State-of-the-Arts

Quantitative evaluation Tables 1, and 2 show the segmentation performance of OGP-Net and other state-of-the-art models for both the settings. The performance of the baseline DeepLabV3+ model for Oracle settings is shown, where it is trained and tested solely on RGB or IR data. To evaluate performance when one modality is missing, all models are trained using both modalities but tested only on IR data as shown in *missing-modality (IR only)*. Segmentation performance for multi-modal fusion is presented in the *Multi-modal Settings*, where OGP-Net consistently outperforms state-of-the-art models with significantly fewer parameters. It outperforms CDDFuse by 4.22 % on average of all datasets. It also outperforms the C.E.N model by 4.94 % on average across all datasets, with only 14.37 % of its total number of parameters.

Qualitative evaluation Fig. 3 shows the predictions of contemporary models and OGP-Net. OGP-Net obtains domain-invariant features across various imaging spectra while preserving modality-specific features for the same object categories. This contributes to the model's capability to make superior predictions. T-SNE feature embeddings for both scenarios are provided in the supplementary material. These embeddings reveal that OGP-Net generates well-structured clusters in the semantic feature space, whereas the baseline

Settings	Method	FMB (mIoU)
Baseline	RGB	49.72
	IR	44.7
RGB-IR	TarDAL (Liu et al. 2022)	50.38
	CDDFuse (Zhao et al. 2023a)	48.41
	DDFM (Zhao et al. 2023b)	35.69
	MFNet (Ha et al. 2017)	39.7
	C.E.N (Wang et al. 2022)	49.01
	MMANet (Wei, Luo, and Luo 2023)	53.97
	OGP-Net (Ours)	**54.89**
IR only	DisOptNet (Kang et al. 2022)	46.34
	MMANet (Wei, Luo, and Luo 2023)	48.59
	RGB2TIR (Lee et al. 2023)	34.7
	ASAPNet (Shaham et al. 2021)	43.1
	TIR ControlNet (Mayr et al. 2024)	47.8
	OGP-Net-IR (Ours)	48.56
	OGP-Net (Ours)	**48.89**

Table 2: Performance comparison of proposed OGP-Net with state-of-the art models on the FMB dataset. Multi-modal settings are denoted as RGB-IR, and IR denotes inference using only IR data.

Method	mIoU
IR (Chen et al. 2018)	61.96
DeepLabV3 + KL	58.5
DeepLabV3 + CWD (Shu et al. 2021)	59.36
DeepLabV3 + DIST (Huang et al. 2022)	61.84
OGP-Net-IR (Ours)	66.73
OGP-Net (Ours)	**66.77**

Table 3: Performance comparison of state-of-the-art knowledge distillation methods with OGP-Net for missing modality scenarios on MSRS dataset. KL refers to Kullback–Leibler divergence.

model produces overlapping clusters.

Ablation Studies

Comparison with distillation techniques Table 3 compares OG-Net and OG-Net-IR with distillation methods, showing that forced feature alignment in KD techniques results in lower performance than the baseline IR model.

Philosophy of feature exchange Different configurations of shared and distinct convolutional and batch normalization layers are verified in Table 4. There is a clear performance decrease when both convolutional and batch normalization layers are shared in the encoder.

Effectiveness of proposed framework Table 5 highlights the effectiveness of OGP-Net components. Multi-view feature maps enhance pixel-level contrastive learning and distillation, with the combination of all techniques yielding the best performance. OGP-Net leverages knowledge from the pre-trained RGB model and achieves rapid convergence in both multi-modal and missing modality scenarios, as illustrated in Figure 4.

Convs	BN	Exchange	mIoU	Param (M)
Unshared	Unshared	✗	69.81	24.89
Shared	Shared	✗	67.23	11.68
Shared	Unshared	✗	69.52	14.24
Shared	Unshared	✓	**69.97**	14.24

Table 4: Ablation experimental results of different versions of OGP-Net on MSRS dataset for multi-modal setting. Exchange refers to the feature exchange strategy. (All results are generated with the EfficientNet-B3 backbone.)

Additional Experiments Additional experiments, including: (i) validating OGP-Net's effectiveness with limited training samples and (ii) comparing performance with MMIF models in missing modality scenarios, are provided in the supplementary material.

Method	mIoU
OGP-Net	**69.97**
OGP-Net - Pixel-wise Contrastive learning	68.98
OGP-Net - GSU	69.62
OGP-Net - Feature Exchange	69.45
OGP-Net - Pixel-wise Contrastive learning - Feature Exchange	69.04

Table 5: Ablation results for contrastive learning, Mixed Feature Exchange and GSU on OGP-Net for multi-modal settings on the MSRS dataset.

Conclusions

This paper introduces OGP-Net, a novel multi-modal fusion approach for multi-spectral semantic segmentation tasks, addressing both multi-modal and missing modality scenarios. It introduces Distillation with Multi-View Contrastive (DMC) and Distillation for Uni-modal Retention (DUR) to preserve both intra- and inter-modal semantic knowledge. DMC aligns unimodal features in a unified latent space, while pixel-level multi-view contrastive learning ensures modality-invariant representations. DUR focuses on retaining modality-specific details within the latent space. OGP-Net consistently delivers superior performance across three public benchmarking datasets, outperforming CDDFuse by an average of **4.74%** in multi-modal settings and surpassing MMANet by **2.74%** in multi-modal and **2.02%** in missing modality settings, all with significantly fewer parameters. Additionally, OGP-Net converges faster and performs better with limited training data.

Acknowledgements

This work was supported by the Centre for Airborne Systems (CABS).

References

Arevalo, J.; Solorio, T.; Montes-y Gómez, M.; and González, F. A. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

Berg, P.; Uzun, B.; Pham, M.-T.; and Courty, N. 2024. Multimodal supervised contrastive learning in remote sensing downstream tasks. *IEEE Geoscience and Remote Sensing Letters*.

Chang, W.-G.; You, T.; Seo, S.; Kwak, S.; and Han, B. 2019. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 7354–7362.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chen, M.; Zheng, Z.; Yang, Y.; and Chua, T.-S. 2023. Pipa: Pixel-and patch-wise self-supervised learning for domain adaptative semantic segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1905–1914.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

De Vries, H.; Strub, F.; Mary, J.; Larochelle, H.; Pietquin, O.; and Courville, A. C. 2017. Modulating early visual processing by language. *Advances in neural information processing systems*, 30.

Do, D. P.; Kim, T.; Na, J.; Kim, J.; Lee, K.; Cho, K.; and Hwang, W. 2024. D3T: Distinctive Dual-Domain Teacher Zigzagging Across RGB-Thermal Gap for Domain-Adaptive Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23313–23322.

Gade, R.; and Moeslund, T. B. 2014. Thermal cameras and applications: a survey. *Machine vision and applications*, 25: 245–262.

Gutmann, M.; and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 297–304. JMLR Workshop and Conference Proceedings.

Ha, Q.; Watanabe, K.; Karasawa, T.; Ushiku, Y.; and Harada, T. 2017. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5108–5115. IEEE.

He, Q. 2024. Prompting multi-modal image segmentation with semantic grouping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2094–2102.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Huang, J.; and Guo, Z. 2023. Pixel-Wise Contrastive Distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16359–16369.

Huang, T.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2022. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35: 33716–33727.

Jain, U.; Wilson, A.; and Gulshan, V. 2022. Multimodal contrastive learning for remote sensing tasks. *arXiv preprint arXiv:2209.02329*.

Ji, W.; Li, J.; Bian, C.; Zhou, Z.; Zhao, J.; Yuille, A. L.; and Cheng, L. 2023. Multispectral Video Semantic Segmentation: A Benchmark Dataset and Baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1094–1104.

Kang, J.; Wang, Z.; Zhu, R.; Xia, J.; Sun, X.; Fernandez-Beltran, R.; and Plaza, A. 2022. DisOptNet: Distilling semantic knowledge from optical images for weather-independent building segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–15.

Kütük, Z.; and Algan, G. 2022. Semantic segmentation for thermal images: A comparative survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 286–295.

Lee, D.-G.; Jeon, M.-H.; Cho, Y.; and Kim, A. 2023. Edge-guided multi-domain rgb-to-tir image translation for training vision tasks with challenging labels. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 8291–8298. IEEE.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

Liang, P.; Jiang, J.; Liu, X.; and Ma, J. 2022. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *European Conference on Computer Vision*, 719–735. Springer.

Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; and Luo, Z. 2022. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5802–5811.

Liu, J.; Liu, Z.; Wu, G.; Ma, L.; Liu, R.; Zhong, W.; Luo, Z.; and Fan, X. 2023. Multi-interactive Feature Learning and a Full-time Multi-modality Benchmark for Image Fusion and Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8115–8124.

Liu, Z.; Li, J.; Shen, Z.; Huang, G.; Yan, S.; and Zhang, C. 2017. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE international conference on computer vision*, 2736–2744.

Mayr, C.; Kubler, C.; Haala, N.; and Teutsch, M. 2024. Narrowing the Synthetic-to-Real Gap for Thermal Infrared Semantic Image Segmentation Using Diffusion-based Conditional Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3131–3141.

Qingyun, F.; Dapeng, H.; and Zhaokui, W. 2021. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*.

Sener, O.; and Koltun, V. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31.

Shaham, T. R.; Gharbi, M.; Zhang, R.; Shechtman, E.; and Michaeli, T. 2021. Spatially-adaptive pixelwise networks for fast image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14882–14891.

Shao, W.; Tang, S.; Pan, X.; Tan, P.; Wang, X.; and Luo, P. 2020. Channel equilibrium networks for learning deep representation. In *International conference on machine learning*, 8645–8654. PMLR.

Shu, C.; Liu, Y.; Gao, J.; Yan, Z.; and Shen, C. 2021. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5311–5320.

Sikdar, A.; Teotia, J.; and Sundaram, S. 2024. SKD-Net: Spectral-based Knowledge Distillation in Low-Light Thermal Imagery for robotic perception. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 9041–9047.

Sikdar, A.; Udupa, S.; Gurunath, P.; and Sundaram, S. 2023. Deepmao: Deep multi-scale aware overcomplete network for building segmentation in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 487–496.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Tang, L.; Yuan, J.; Zhang, H.; Jiang, X.; and Ma, J. 2022. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83: 79–92.

Tian, Z.; Chen, P.; Lai, X.; Jiang, L.; Liu, S.; Zhao, H.; Yu, B.; Yang, M.-C.; and Jia, J. 2022. Adaptive perspective distillation for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 1372–1387.

Udupa, S.; Gurunath, P.; Sikdar, A.; and Sundaram, S. 2024. MRFP: Learning Generalizable Semantic Segmentation from Sim-2-Real with Multi-Resolution Feature Perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5904–5914.

Valada, A.; Mohan, R.; and Burgard, W. 2020. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision*, 128(5): 1239–1285.

Wang, Y.; Sun, F.; Huang, W.; He, F.; and Tao, D. 2022. Channel exchanging networks for multimodal and multitask dense image prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5481–5496.

Wang, Y.; Sun, F.; Lu, M.; and Yao, A. 2020. Learning deep multimodal feature representation with asymmetric multi-layer fusion. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3902–3910.

Wei, S.; Luo, C.; and Luo, Y. 2023. MMANet: Margin-aware Distillation and Modality-aware Regularization for Incomplete Multimodal Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20039–20049.

Yang, C.; Zhou, H.; An, Z.; Jiang, X.; Xu, Y.; and Zhang, Q. 2022. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12319–12328.

Ye, J.; Lu, X.; Lin, Z.; and Wang, J. Z. 2018. Rethinking the smaller-norm-less-informative assumption in channel pruning of convolution layers. *arXiv preprint arXiv:1802.00124*.

Yu, Y.; Xiong, Y.; Huang, W.; and Scott, M. R. 2020. Deformable siamese attention networks for visual object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6728–6737.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, 818–833. Springer.

Zhang, J.; Liu, R.; Shi, H.; Yang, K.; Reiß, S.; Peng, K.; Fu, H.; Wang, K.; and Stiefelhagen, R. 2023. Delivering Arbitrary-Modal Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1136–1147.

Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6848–6856.

Zhao, D.; Yuan, B.; and Shi, Z. 2023. Inherit with distillation and evolve with contrast: Exploring class incremental semantic segmentation without exemplar memory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 11932–11947.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; and Van Gool, L. 2023a. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5906–5916.

Zhao, Z.; Bai, H.; Zhu, Y.; Zhang, J.; Xu, S.; Zhang, Y.; Zhang, K.; Meng, D.; Timofte, R.; and Van Gool, L. 2023b. DDFM: denoising diffusion model for multi-modality image fusion. *arXiv preprint arXiv:2303.06840*.

Zheng, Z.; Ma, A.; Zhang, L.; and Zhong, Y. 2021. Deep multisensor learning for missing-modality all-weather mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 174: 254–264.