2024 IEEE International Conference on Robotics and Automation (ICRA)
May 13-17, 2024. Yokohama, Japan

2024 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-8457-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRA57147.2024.10611323

SKD-Net: Spectral-based Knowledge Distillation in Low-Light Thermal Imagery for robotic perception

Aniruddh Sikdar*[1], Jayant Teotia*[2] and Suresh Sundaram[3]

***Abstract*— Enhancing the generalization capacity for semantic segmentation of aerial perception systems for safety-critical applications is vital, especially for environments with low-light and adverse conditions. Multi-spectral fusion techniques aim to maintain the merits of electro-optical (EO) and infrared (IR) images, e.g., retaining low-level features and capturing detailed textures from both modalities. However, these techniques encounter limitations when faced with scenarios involving missing modalities, especially during inference when only IR images are available. In this paper, we propose a novel spectral-based knowledge distillation architecture known as SKD-Net to improve the performance of deep learning models for missing modality scenarios for semantic segmentation tasks. In this architecture, we make use of Gated Spectral Unit to combine information from both modalities. SKD-Net aims to extract valuable semantic information from EO images while preserving spectral knowledge from the IR images within the feature space. The model retains the style information in the shallow layers while simultaneously fusing the high-level semantic context obtained from EO and IR images to improve the feature generation capacity when dealing with only IR images during inference. SKD-Net outperforms state-of-the-art multi-modal fusion and distillation models by 2.8% on average in scenarios with missing modalities when using only IR data during inference in two public benchmarking datasets. This performance increase is achieved without additional computational costs compared to the baseline segmentation models.**

I. INTRODUCTION

Aerial perceptual robustness plays an important role in enabling UAVs to operate effectively across diverse environments, including those with harsh conditions and low illumination. This is valuable while conducting inspections [32], searches [33], [35], and surveillance operations [34] in challenging environments using autonomous drones. A lot of research focused on achieving high segmentation performance, but a gap remains for enhancing generalizing capabilities, particularly in adverse environmental conditions. The majority of vision models have been developed with a focus on cameras operating in the visible spectrum [36], primarily because of the ready availability of large-scale RGB datasets [16],[17]. However, these models often experience a drop in performance when operating in low-light conditions and adverse environments. Since IR waves carry distinct spectral information and have the capability to penetrate dust and smoke, IR cameras can be used in challenging environments with limited visibility and low-light conditions. Hence, an increase in segmentation performance can be achieved in adverse conditions.

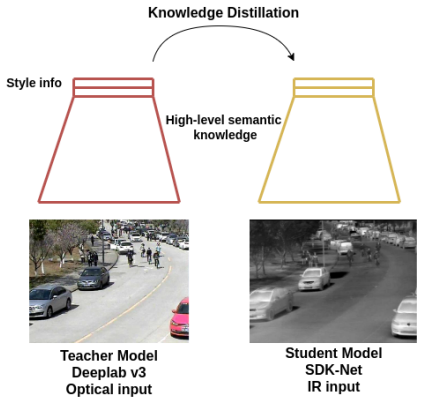

Fig. 1. Proposed knowledge distillation framework encourages SKD-Net to transfer semantic knowledge between optical and IR modalities.

Despite the utility of IR cameras, the images generated contain less semantic information in comparison to EO images, resulting in significant drops in the performance of deep learning models for the dense prediction task. This leads to the requirement for multi-sensor fusion approaches that combine both EO and IR images addressing the limitations of individual modalities. One popular approach involves utilizing multi-spectral fusion strategies that incorporate EO and IR images. These methods have garnered attention, particularly with the accessibility of co-registered EO-IR datasets [1][2][18]. Multi-spectral networks can be trained on paired RGB-thermal image datasets for feature extraction and image fusion, like Cross-modality transformer [19], GAFF [20], and CDDFuse [21]. These fusion approaches improve the overall urban semantic segmentation performance by combining both modalities, but they often rely on co-registered images during both, training and testing. There are not many publicly available coregistered datasets and collecting a custom dataset can be quite costly [22]. Additionally, when data from one of the sensors becomes corrupted, it can substantially degrade the performance of these models. In practical application scenarios, a major challenge is the unavailability of all data sources consistently. This is commonly called the ”missing modality problem” [24].

*Equal contribution of both the authors.

This work was supported by the Ministry of Electronics and Information Technology, India.

[1]Aniruddh Sikdar is with the Robert Bosch Centre for Cyber-Physical Systems, Indian Institute of Science, Bangalore. aniruddhss@iisc.ac.in

[2]Jayant Teotia is with Robert Bosch Centre for Cyber-Physical Systems, Indian Institute of Science, Bangalore. jayantteotia@iisc.ac.in

[3]Suresh Sundaram is with the Department of Aerospace Engineering, Indian Institute of Science, Bangalore. vssuresh@iisc.ac.in

When using the multisensor information to train an existing model, typical approaches rely on two key methods: knowledge transfer [23] and knowledge distillation [25]. The absence of a knowledge retention mechanism can result in the loss of information during the transfer process [25]. Knowledge distillation [9] can be used to distill the knowledge, but directly matching these images and the distributions of modality-specific features with large domain gaps can result in negative transfer, primarily due to the forced feature alignment.

This paper proposes a novel spectral-based knowledge distillation architecture known as SKD-Net to address the missing modality scenario as shown in Fig. 1. In contrast to methods that rely on image fusion, our approach centers on the fusion of image modalities at the feature map level. The aim is to extract the spectral knowledge pertaining to the same object across different modalities and to learn the domain invariant and domain-specific feature representation from cross-modal data, especially the semantic context of the objects. A novel module called Gated Spectral Fusion is proposed to combine the spectral information from multiple imaging modalities for efficient knowledge distillation. The main contributions are as follows:

- A novel spectral-based knowledge distillation architecture known as SKD-Net is proposed for the missing modality scenario. SKD-Net helps in narrowing the segmentation performance difference between IR and EO images.
- The proposed SKD-Net consists of shared encoders for both EO and IR, trained using contrastive loss for intra-class compactness while retaining the style information. The model also distills multi-level semantic features of optical images to fuse the rich semantic information.
- A feature reuse strategy is adopted to avoid additional computational costs. This results in the same computation complexity as the baseline segmentation model, with increased performance for IR images.
- The proposed SKD-Net model outperforms other state-of-the-art multi-modal fusion models by an average of 2.8% on two publicly available benchmarking datasets.

II. Related Work

A. Knowledge Distillation

Knowledge distillation [9] originally aimed to transfer knowledge from a complex network to a smaller one by reducing the classification gap using soft targets. Subsequently, the pixel-level knowledge distillation strategy has gained significant attention for training compact models in the context of semantic segmentation tasks. Balancing the trade-off between high accuracy and high speed remains a persistent challenge in the field of semantic image segmentation. A knowledge distillation framework named double similarity distillation (DSD) has been introduced to enhance classification accuracy by capturing similarity knowledge in both pixel and category dimensions. Additionally, a Pixelwise Similarity Distillation (PSD) module is presented to capture finer spatial dependencies, as proposed in [10]. Distillation techniques have also been explored to address the challenge of handling missing modalities [11], [12]. DisOptNet [3] introduced a distillation technique aimed at transferring semantic knowledge from the optical to SAR (Synthetic Aperture Radar) modality. This approach is designed to enhance segmentation performance, particularly in scenarios where one modality is missing, as applied in weather-independent urban mapping applications.

B. Contrastive Learning

Contrastive learning has been widely used for the purpose of learning representations in the absence of labeled data [26], and it has demonstrated significant superiority over other pretext task-based options. Recent works [27] show the use of label information for image-level pretraining. Contrastive learning aims to facilitate the learning of discriminative feature representations by distinguishing between similar feature pairs and dissimilar (negative) pairs. The positive pair sampling strategy involves applying strong perturbations to generate diverse views [28]. Negative pairs, on the other hand, can be generated through random sampling or more advanced techniques like negative mining [29]. PiPa [15] has been proposed for unsupervised domain adaptation to facilitate intra-image pixel-wise correlations and patch-wise semantic consistency against different contexts to promote intraclass compactness and inter-class separability. A pixel-wise training strategy utilizing contrastive learning has been proposed [30] for an inter-image, pixel-to-pixel paradigm that uses the global semantics of labeled pixels for supervised learning. The correlation between individual pixels and pixel and semantic regions is optimized. One of the pioneering works using contrastive learning for knowledge distillation [31], used contrastive-based objective. This objective function encourages the teacher and student models to map the same input to similar representations.

C. Multi-modal Image Fusion

Deep learning has been used for a variety of vision applications, with an increasing focus on learning for different modalities [37][38]. Multi-modal fusion techniques fuse the data from different modalities to capture cross-modality features. The Cross-modality transformer network [19] is designed to acquire long-range dependencies and use global contextual information during the feature extraction phase. In CDDFuse [21], Restormer blocks are employed to extract shallow features across different modalities, and a dual-branch Transformer is used to incorporate long-range attention mechanisms for handling low-frequency global features. Additionally, Invertible Neural Network (INN) blocks are incorporated to extract high-frequency local information. [25] showed that the distinction among deep models trained on data from diverse modalities can be attributed to the parameter distribution of sensor-invariant and sensor-specific operations. A prototype network was introduced to acquire meta-sensory representation by modeling the mechanism to

retain knowledge using an alignment operation. Channel-Exchanging-Network (CEN) [4] proposed a dynamic swapping of channels between sub-networks as a mechanism for fusing information from various modalities. The process is self-directed and relies on assessing the importance of individual channels by evaluating the magnitude of the Batch-Normalization (BN) scaling factor during the training process.

III. SKD-Net: Spectral-based Knowledge Distillation Network

A. Problem Formulation

Thermal imagery holds significance in low-light scenarios, however, deep-learning models experience a drop in performance when trained on IR images, mainly because they contain fewer semantic information than EO images. The aim is to enhance the model's representation ability across various modalities during training and to maintain this knowledge when performing inference with only one modality. Let $\{X^O, X^I\}$ = $\{(X_1^O, X_1^I)$, ..., $(X_n^O, X_n^I)\}$ denote the co-registered EO-IR image pairs from dataset *D*, with their corresponding pixel-wise labels Y = $\{Y_1, ..Y_n\}$. Images from both modalities, i.e., $\{X^O, X^I\}$, are passed as input to the semantic segmentation model f_θ during the training process, where θ represents the learnable parameters. However, only the IR images, represented as X^I, are passed to the model during inference. Spectral-based knowledge distillation is proposed using SKD-Net to acquire domain-invariant features from both modalities.

B. SKD-Net Architecture

Semantic segmentation networks commonly employ an encoder-decoder architecture and can be denoted as f_θ, where θ represents the learnable parameters. The i^{th} encoding stage is denoted as $f_i(.)$, and maps the features F_{i-1} to F_i, and $d(.)$ denotes the decoder layers, as shown in Fig. 2. Since the shallow layers in CNNs retain the style-related information by capturing local structures [13], the first three encoder blocks from the backbones of f_θ are shared and are represented as $\mathbf{F}_3$. The features become more domain-specific in the later layers of the model. Consequently, these layers are retained individually to preserve semantic information primarily derived from EO images. Hence, $\mathbf{F}_3$ is passed to the EO and IR branches to capture domain-specific knowledge. The outputs of the decoder for both the EO and IR branches are represented as $\mathbf{F}_I$ and $\mathbf{F}_O$, and are passed to the Gated Spectral Unit (GSU) block as shown in Fig. 2 (b).

GSU is proposed to enforce spectral learning, inspired by gated multimodal units [14]. The main idea for the multiplicative gates is to determine which input has a greater impact on generating the correct output for a rich multimodal representation. This approach avoids manual adjustments and enables the model to learn from the training data independently. It helps to learn the spectral properties from the EO and IR branches and learns to decide the influence of different units' activation using gates. Fig. 2 (b) depicts the structure of a GSU. The output of the EO and IR branches and their summation are passed to the GSU block. These outputs are passed through convolution layers and then with the $tanh$ activation function, as given below,

$$h_1 = tanh(W_1 * F_I) \tag{1}$$

$$h_2 = tanh(W_2 * F_O) \tag{2}$$

$$h_3 = tanh(W_3 * (F_I + F_O)) \tag{3}$$

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

$$Z_1 = \sigma(W_1 \otimes [F_I, F_O, F_I + F_O]) \tag{4}$$

where $[\cdot, \cdot]$ denotes the concatenation operator and σ denotes sigmoid operation. The final output predictions of the fusion block F^f is given by,

$$F^f = Z_1 \otimes h_1 + Z_2 \otimes h_2 + Z_3 \otimes h_3 \tag{5}$$

where $\otimes$ represents the multiplication operation. SKD-Net model has three outputs as shown in the figure, two for each modality, i.e., for optical and IR, and the third output from the GSU block, given by F^f. GSU is used for distillation only and is removed during inference.

C. Training Scheme

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

1) Training Step 1: The baseline segmentation model is trained with the optical images with their corresponding labels $\{X^O, Y\}$, using the $L_{ST1}(p, y)$ consisting of contrastive loss for encoder layers, and segmentation loss to train the whole model. The segmentation loss $L_{seg}(p, y)$ consisting of summation of cross-entropy and dice loss, given by,

$$L_{seg}(p, y) = -\sum_i y_i log(p_i) + 1 - \frac{2\sum_i p_i y_i}{\sum_i y_i + \sum_i p_i} \tag{6}$$

where, *y* and *p* denote the ground truth labels and the pixel-wise predictions respectively.

To train the encoder for superior style representation, contrastive learning L_{CL} is employed in the feature space, which is taken from the first two layers of the encoder to improve the intra-domain mining for optical modality. This involves mapping image pixels into an embedding space using a projection head h_{pixel}, facilitating discriminative feature learning. This process aims to bring pixel embeddings from the same category closer together while pushing pixel embeddings from different categories farther apart. In doing so, the model is encouraged to learn correlations between labeled pixels. Using the pixel-wise labels, pixels belonging to the same class are treated as positive samples, while

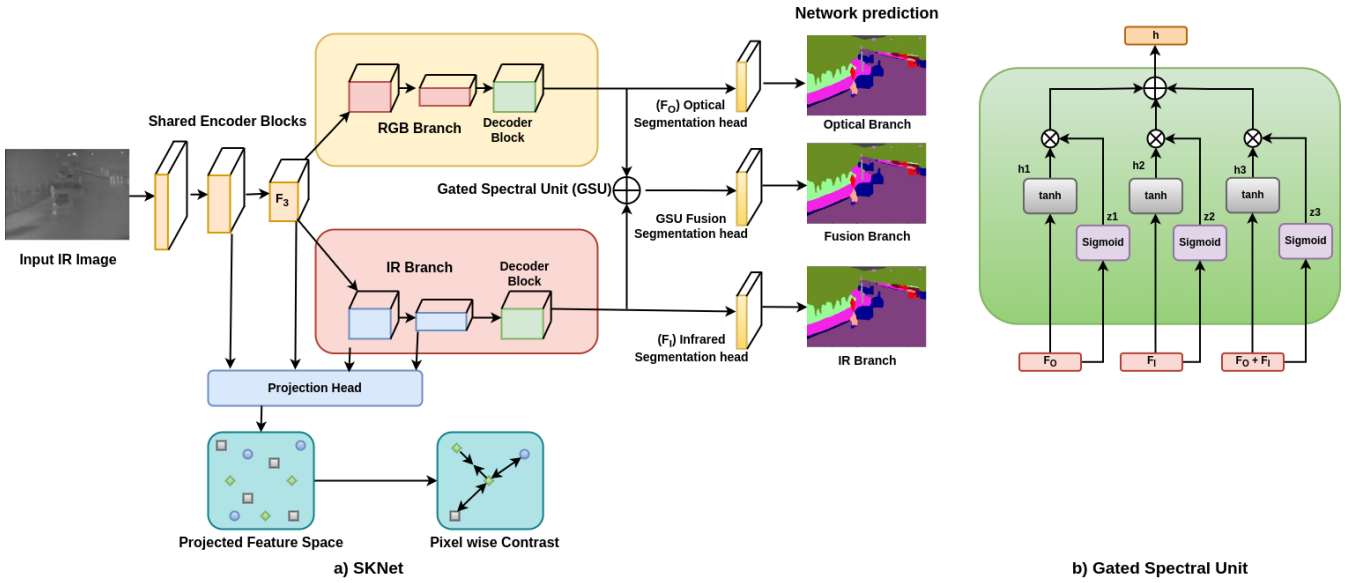

Fig. 2. Spectral-based knowledge distillation network (SKD-Net) architecture. a) The first three encoder blocks are shared between EO and IR branches. The output of the decoder block of the two branches is fed into GSU for feature fusion. Pixel-wise contrastive loss is used for features from the four encoder blocks (two from shared and two from the IR branch). b) Gated Spectral Unit takes three inputs from the EO branch, IR branch, and pixel-wise addition of the features of the two branches and outputs a single feature map.

those belonging to different classes are considered negative samples. The pixel-wise contrastive loss is formulated as,

$$L_{CL} = -\sum_{C(i)=C(j)} log \frac{r(e_i, e_j)}{\sum_{k=1}^{Np} r(e_i, e_j)} \quad (6)$$

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. The similarity is calculated using the exponential similarity function: r(e_i, e_j) = exp($s(e_i, e_j)$ / τ), where s represents the cosine similarity, and τ is the temperature parameter. A semi-hard example sampling strategy [15] is adopted, where the negative samples are retained from the whole training batch, with the top 10% nearest negatives and farthest positives selected for each anchor sampling.

2) Training Step 2: The SKD-Net architecture is trained, and distillation from the pre-trained optical model is performed using two distillation loss terms, namely L_{D1} and L_{D2}. Using the multi-class pixel-wise predictions obtained by the pre-trained model denoted as p^{PO}, the distillation loss L_{D1} is given by,

$$L_{D1}(p, p^{PO}) = \sum_i p_i^{PO} log \frac{p_i^{PO}}{p_i} - \sum_i p_i^{PO} log(p_i) \quad (7)$$

where *p* represents the predictions made by the EO branch of SKD-Net. Kullback-Leiber (KL) divergence term along with the cross-entropy loss is used to generate similar predictions made by the pre-trained optical branch. Adopting the deep distillation strategy from DisOptNet [3], the multi-level semantic information is distilled from the EO branch of the pre-trained model to the EO branch of SKD-Net, using the mean square error loss, given by,

$$L_{D2}(F, F^{PO}) = \sum_{i \in \{4,5\}} \|F_i - F_i^{PO}\|_2 + \|F_d - F_d^{PO}\|_2 \quad (8)$$

which measures the difference between the features of the last two layers of the encoders and the decoder output. The joint loss function for training the SKD-Net is given by,

$$L_{ST2} = L_{seg}(y, p) + L_{D1}(p, p^{PO}) + L_{D2}(F, F^{PO}) + L_{CL} \quad (9)$$

where the segmentation loss L_{seg} is used to optimize the model's output from fused segmentation and the IR segmentation head, along with contrastive loss L_{CL}. In the second training step, the last four features of the encoders are passed to the projection head and L_{CL} is used to improve the intra-domain mining for IR modality, which is helpful when performing inference on IR modality only. During inference, only the IR branch of SKD-Net is used, which has the same configuration as DeepLabV3+.

IV. Experimental Results

A. Experimental Setup

The performance of the proposed SKD-Net is compared with two existing methods: DisOptNet [3] and Multimodal Fusion by Channel Exchanging (C.E.N) [4]. During training, the models are trained using both optical and Infrared (IR) modalities, however during inference, only IR images are accessible. This requires the evaluation of these models in missing modality scenarios. All the models are evaluated using two datasets, MSRS [1] and MVSS [2]. To ensure fair comparisons, the baseline model DeepLabV3+, and DisOptNet are re-implemented using the same training strategy as SKD-Net. Additionally, C.E.N is re-implemented

TABLE I

PERFORMANCE COMPARISON OF IOU (%) OF SKD-NET WITH OTHER STATE-OF-THE-ART MODELS ON MSRS DATASET.

Method	Publication	mIoU	# of Params
Baseline (VI)		64.33	11.68 M
Baseline (IR)		61.7	11.68 M
DisOptNet	TGRS 2022	63.38	11.68 M
C.E.N	TPAMI 2022	62.93	99.13 M
SKD-Net		**64.67**	11.68 M

TABLE II

PERFORMANCE COMPARISON OF IOU (%) OF SKD-NET WITH OTHER STATE-OF-THE-ART MODELS ON MVSS DATASET.

Method	Publication	mIoU	# of Params
Baseline (VI)		48.55	11.68 M
Baseline (IR)		42.82	11.68 M
DisOptNet	TGRS 2022	43.22	11.68 M
C.E.N	TPAMI 2022	40.33	99.13 M
SKD-Net		**45.53**	11.68 M

with the training strategy specified in [4]. The evaluation of segmentation performance is based on the Intersection over Union (IoU) metric.

1) Datasets Description: The experiments are conducted on two semantic segmentation datasets, namely MSRS and MVSS, which consist of aligned visible and infrared images. The MSRS dataset [1], comprises 1,444 high-quality image pairs with a spatial resolution of 480 × 640, each accompanied by pixel-wise semantic labels. It contains 9 classes, including categories such as color cones, cars, bikes, and pedestrians. The dataset is randomly partitioned into training and test samples, with 1083 samples allocated for training and 361 samples for inference. The MVSS dataset [2] contains a diverse array of urban scenes, encompassing both daytime and nighttime conditions, presenting a range of challenges. It has a total of 1616 samples, with 1004 training samples, and 612 samples for testing. The image has been resized to 320x480 as specified in MVSS [2] as the original images are of dis-similar dimensions. The dataset contains 26 classes, such as cars, buses, motorcycles, poles, buildings, and pedestrians.

2) Implementation Details: The baseline segmentation model used is DeepLabV3+ [5] with EfficientNet-B3 [8] backbone, pre-trained on ImageNet [7]. Data augmentations for both the training steps include horizontal flips with 50% probability. In all experiments, the models are trained using the SDG optimizer, starting with an initial learning rate of 5 $\times$ 10^{-3}. A polynomial scheduler decreases the learning rate after each epoch, a decay factor of (1 - step/total steps)$^{0.9}$.

TABLE III

ABLATION ANALYSIS WITH SETTING OF SKD-NET.

Method	mIoU
Baseline	42.82
DisOptNet	43.22
SKD-Net w/o Contrastive loss	44.50
SKD-Net w/o GSU	45.33
SKD-Net	45.53

The models are trained with a batch size of 8, for 200 epochs. All experiments are conducted using an NVIDIA Quadro RTX 5000 GPU.

B. Quantitative Evaluation

Table I and Table II show the segmentation performance of SKD-Net and other state-of-the-art multimodal fusion models. Both tables show the performance of the baseline DeepLabV3+ model for oracle settings, where it is trained and tested solely on visible (VI) or infrared (IR) data. To evaluate the performance when one modality is missing, all other models are trained using pairs of electro-optical (EO) and infrared (IR) data but are exclusively tested on IR data. As shown in Table I, SKD-Net exhibits superior performance for missing modality scenarios for the MSRS dataset. It demonstrates a 2.97% improvement compared to the baseline model. Additionally, SKD-Net outperforms DisOptNet by 1.29% and C.E.N. by 1.75%. It also outperforms the baseline model trained and evaluated exclusively on optical images, demonstrating the effectiveness of the knowledge distillation technique of SKD-Net. Table II presents the performance results for the missing modality scenario on the MVSS dataset. Since only the baseline DeepLabV3+ model is used for inference, there is no increase in the model complexity. It outperforms baseline and DisOptNet by 2.71% and 2.31% respectively. SKD-Net consistently outperforms C.E.N model, with only 12% of its total number of parameters.

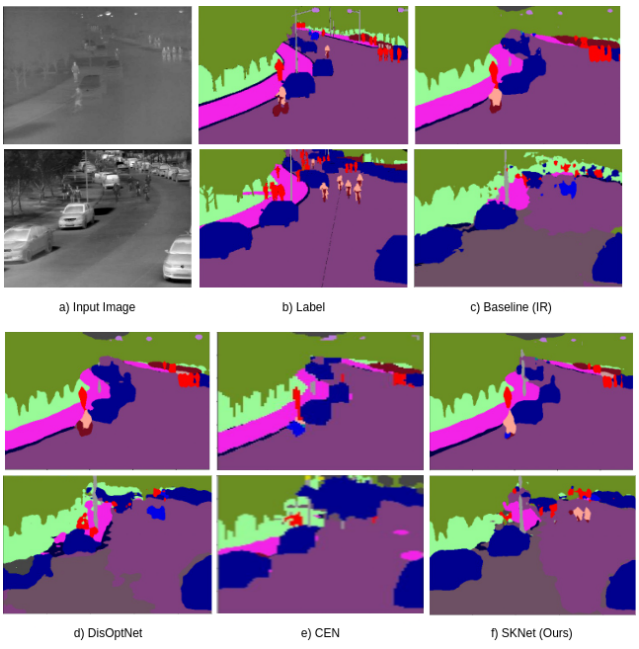

Fig. 3. Comparison of output predictions of SKD-Net with baseline and state-of-the-art models. The cluttered car labels (blue) can be seen in (c-e), as opposed to the ground truth labels. SKD-Net(f) is able to segment cars more accurately. Our model is also able to predict bicycles (light brown) in the second image, as opposed to the other models.

C. Qualitative Evaluation

The output predictions of state-of-the-art models trained on MVSS dataset are shown in Fig. 3. SKD-Net performs

feature distillation from optical images and acquires domain-invariant features across various spectra for the same object categories. This contributes to the model's ability to make superior predictions, particularly in roads, vehicles, and pedestrians, where its performance surpasses that of other models. In the figure, it can be seen that SKD-Net predicts bicycles and street lights while the other models are not able to predict them. In SKD-Net, contrastive learning helps to better distinguish objects from each other. This can be seen in the predictions of car for both the images, while the other models give overlapping predictions.

D. Ablation Study

An ablation study was conducted, as shown in Table III, to evaluate the significance of individual components within SKD-Net, involving the removal of both the GSU block and the contrastive learning loss. All experiments conducted have the same training settings mentioned in the implementation details. The table shows that the GSU gives a performance boost of 1.68% as compared to the baseline model, by aiding the knowledge distillation of the domain invariant features across diverse modalities. The utilization of contrastive learning assists in preserving the style information during the distillation process and improves the performance by 2.49% as compared to the baseline. Combining both results in the best overall performance.

V. CONCLUSIONS

This paper introduces a novel spectral-based knowledge distillation scheme known as SKD-Net for semantic segmentation tasks, for missing modality scenarios. The training of the encoders involves using contrastive loss to preserve intra-modality knowledge and a novel module known as Gated Spectral Fusion is also proposed to aid this distillation process. SKD-Net consistently achieves superior performance on two public benchmarking datasets. It has an improvement of 2.8% on average compared to the baseline segmentation model without any increase in the model complexity during the inference phase. It also outperforms C.E.N by 3.2% on average with only 12% of its parameters.

REFERENCES

[1] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. Inf. Fusion, 83-84:79–92, 2022.

[2] 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 (pp. 1094-1104).

[3] 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, pp.1-15.

[4] Wang, Y., Huang, W., Sun, F., Xu, T., Rong, Y. and Huang, J., 2020. Deep multimodal fusion by channel exchanging. Advances in neural information processing systems, 33, pp.4835-4845.

[5] Jiang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, Feb. 2018

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceed- ings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[8] Tan, M. and Le, Q., 2019, May. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.

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

[10] Feng, Y., Sun, X., Diao, W., Li, J. and Gao, X., 2021. Double similarity distillation for semantic image segmentation. IEEE Transactions on Image Processing, 30, pp.5363-5376.

[11] Garcia, N.C., Morerio, P. and Murino, V., 2018. Modality distillation with multiple stream networks for action recognition. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 103-118).

[12] Crasto, N., Weinzaepfel, P., Alahari, K. and Schmid, C., 2019. Mars: Motion-augmented rgb stream for action recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7882-7891).

[13] Matthew D. Zeiler and Rob Fergus. Visualizing and under- standing convolutional networks. In David Fleet, Tomas Pa- jdla, Bernt Schiele, and Tinne Tuytelaars, editors, Computer Vision – ECCV 2014, pages 818–833, Cham, 2014. Springer International Publishing.

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

[15] Chen, M., Zheng, Z., Yang, Y. and Chua, T.S., 2022. Pipa: Pixel-and patch-wise self-supervised learning for domain adaptative semantic segmentation. arXiv preprint arXiv:2211.07609.

[16] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13 (pp. 740-755). Springer International Publishing.

[17] Cordts, M., Omran, M., Ramos, S., Scharwächter, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B., 2015, June. The cityscapes dataset. In CVPR Workshop on the Future of Datasets in Vision (Vol. 2). sn.

[18] Free teledyne flir thermal dataset for algorithm training, [online] Available: https://www.flir.com/oem/adas/adas-dataset-form/.

[19] F. Qingyun, H. Dapeng and W. Zhaokui, "Cross-modality fusion transformer for multispectral object detection", arXiv preprint, 2021.

[20] H. Zhang, E. Fromont, S. Lefèvre and B. Avignon, "Guided attentive feature fusion for multispectral pedestrian detection", Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 72-80, 2021.

[21] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5906-5916.

[22] Z. Kütük and G. Algan, "Semantic segmentation for thermal images: A comparative survey", Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 286-295, 2022.

[23] Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[24] Garcia, N.C., Morerio, P. and Murino, V., 2018. Modality distillation with multiple stream networks for action recognition. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 103-118).

[25] 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, pp.254-264.

[26] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In ECCV, 2016.

[27] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In CVPR, 2021.

[28] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In ICML, 2020.

[29] Mahmut Kaya and Hasan S¸ akir Bilge. Deep metric learning: A survey. Symmetry, 11(9):1066, 2019.

[30] Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E. and Van Gool, L., 2021. Exploring cross-image pixel contrast for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7303-7313).
[31] Tian, Y., Krishnan, D. and Isola, P., 2019. Contrastive representation distillation. arXiv preprint arXiv:1910.10699.
[32] Senthilnath, J., K. Harikumar, and Suresh Sundaram. "Metacognitive Decision-Making Framework for Multi-UAV Target Search Without Communication." IEEE Transactions on Systems, Man, and Cybernetics: Systems (2024).
[33] Velhal, Shridhar, Suresh Sundaram, and Narasimhan Sundararajan. "Dynamic resource allocation with decentralized multi-task assignment approach for perimeter defense problem." IEEE Transactions on Aerospace and Electronic Systems 58.4 (2022): 3313-3325.
[34] John, Josy, K. Harikumar, J. Senthilnath, and Suresh Sundaram. "An Efficient Approach With Dynamic Multiswarm of UAVs for Forest Firefighting." IEEE Transactions on Systems, Man, and Cybernetics: Systems (2024).
[35] Gurumurthy, Vignesh, Nishant Mohanty, Suresh Sundaram, and Narasimhan Sundararajan. "An efficient reinforcement learning scheme for the confinement escape problem." Applied Soft Computing 152 (2024): 111248.
[36] Sikdar, Aniruddh, Sumanth Udupa, Prajwal Gurunath, and Suresh Sundaram. "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, pp. 487-496. 2023.
[37] Sikdar, Aniruddh, Sumanth Udupa, Suresh Sundaram, and Narasimhan Sundararajan. "Fully Complex-valued Fully Convolutional Multi-feature Fusion Network (FC 2 MFN) for Building Segmentation of InSAR images." In 2022 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 581-587. IEEE, 2022.
[38] Sikdar, Aniruddh, Sumanth Udupa, and Suresh Sundaram. "Fully complex-valued deep learning model for visual perception." In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5. IEEE, 2023.