

Asymmetric Cross-Modal Knowledge Distillation: Bridging Modalities with Weak Semantic Consistency

Riling Wei^{1*}, Kelu Yao^{1,2*}, Chuanguang Yang³, Jin Wang⁴, Zhuoyan Gao¹, Chao Li^{1†}

¹Research Center for Space Computing System, Zhejiang Laboratory, Hangzhou, China

²Zhejiang University, Hangzhou, China

³Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

⁴The University of Hong Kong, Hong Kong, China

{weirl, yaokelu, gaozy, lichao}@zhejianglab.org, yangchuanguang@ict.ac.cn, wj0529@connect.hku.hk

Abstract

Cross-modal Knowledge Distillation has demonstrated promising performance on paired modalities with strong semantic connections, referred to as Symmetric Cross-modal Knowledge Distillation (SCKD). However, implementing SCKD becomes exceedingly constrained in real-world scenarios due to the limited availability of paired modalities. To this end, we investigate a general and effective knowledge learning concept under weak semantic consistency, dubbed Asymmetric Cross-modal Knowledge Distillation (ACKD), aiming to bridge modalities with limited semantic overlap. Nevertheless, the shift from strong to weak semantic consistency improves flexibility but exacerbates challenges in knowledge transmission costs, which we rigorously verified based on optimal transport theory. To mitigate the issue, we further propose a framework, namely SemBridge, integrating a Student-Friendly Matching module and a Semantic-aware Knowledge Alignment module. The former leverages self-supervised learning to acquire semantic-based knowledge and provide personalized instruction for each student sample by dynamically selecting the relevant teacher samples. The latter seeks the optimal transport path by employing Lagrangian optimization. To facilitate the research, we curate a benchmark dataset derived from two modalities, namely Multi-Spectral (MS) and asymmetric RGB images, tailored for remote sensing scene classification. Comprehensive experiments exhibit that our framework achieves state-of-the-art performance compared with 7 existing approaches on 6 different model architectures across various datasets.

Code — <https://github.com/weirl-922/ACKD>

Introduction

Cross-modal Knowledge Distillation (CMKD) (Huo et al. 2024; Wang et al. 2024; Dai, Das, and Bremond 2021; Li et al. 2022; Xue et al. 2022) has demonstrated remarkable performance in various tasks such as visual recognition (Zhao et al. 2024; Lu et al. 2024; Kim et al. 2024) and audio-visual classification (Sarkar and Etemad 2024; Ren

*These authors contributed equally.

†Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



Figure 1: SCKD distills knowledge between modalities from the same location, assuming strict semantic alignment. In contrast, ACKD enables cross-modal transfer with only weak semantic consistency, regardless of location.

et al. 2021), by distilling complementary knowledge across modalities from teacher to student models.

Compared to conventional computer vision tasks, remote sensing (RS) tasks often involve richer and more diverse data modalities, e.g., Multi-Spectral (MS) images (Kettig and Landgrebe 1976), Hyper-Spectral (HS) images (Landgrebe 2002), Light Detection And Ranging (LiDAR) (Reutebuch, Andersen, and McGaughey 2005), etc., making them particularly well-suited for cross-modal learning.

Recently, this potential has attracted increasing attention, and numerous studies have explored the application of CMKD to RS scenarios. In scene classification (Cheng, Han, and Lu 2017), researchers (Liu, Qu, and Zhang 2022; Shin et al. 2023) have used MS images as the teacher modality to distill knowledge into RGB images via CMKD, resulting in significantly improved performance of the RGB-based models. In land cover classification (Phiri and Morgenroth 2017), CMKD is applied to address the issue of missing modalities during inference, and demonstrated its effectiveness across several multi-modal RS datasets (Wang et al. 2023).

Despite its promising potential in RS, CMKD still faces

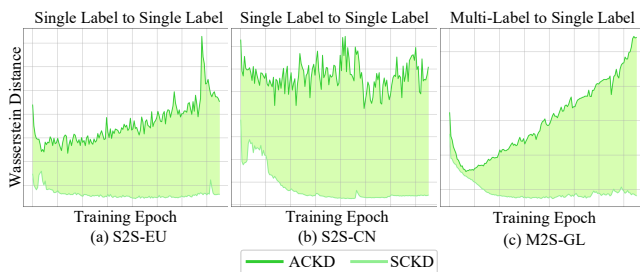


Figure 2: Wasserstein distance between ACKD and SCKD on three datasets. ACKD consistently incurs higher transport costs than SCKD during training, reflecting the challenge of cross-modal alignment in asymmetric settings.

notable challenges in real-world applications. Most existing approaches (Liu, Qu, and Zhang 2022; Shin et al. 2023; Wang et al. 2023) inherently assume that the modalities used by the teacher and student share the same semantic content (*i.e.*, paired data), a setup we collectively refer to as Symmetric Cross-modal Knowledge Distillation (SCKD).

However, in practice, the application of SCKD is often constrained by the scarcity of paired data, primarily due to the quantity imbalance arising from the high acquisition cost of teacher modalities. For example, MS images, which are commonly employed as teacher modalities, typically outperform RGB images in scene-understanding tasks due to their higher spectral resolution (Park et al. 2007; Ma, Yuan, and Kozak 2023). Nevertheless, collecting MS data requires specialized equipment, posing significant challenges for large-scale deployment. In contrast, lower-information-density modalities, such as RGB images, are far more accessible through satellites, UAVs, and other widely available platforms (Liu, Qu, and Zhang 2022; Cheng, Han, and Lu 2017). Consequently, only a small fraction of RGB images are accompanied by corresponding MS modalities, limiting the scalability and practicality of SCKD.

This challenge underscores the need for more flexible distillation strategies that can operate effectively under unpaired or weakly paired settings. A natural and important question thus arises: *Is it possible to distill knowledge between modalities that do not share strong semantic correspondence, such as MS images collected from Europe and RGB images captured in Asia?* We refer to this setting as Asymmetric Cross-modal Knowledge Distillation (ACKD), as illustrated in Figure 1.

Accordingly, ACKD is proposed to overcome the limitations of SCKD in unpaired scenarios by facilitating knowledge transfer between modalities with significant semantic discrepancies. As shown in Table 3 and Table 4, multiple SCKD methods fail to achieve satisfactory performance when applied directly to ACKD. In certain cases, the performance even drops below that of the uni-modal baseline without any distillation, indicating that directly transferring SCKD strategies to asymmetric scenarios is ineffective.

To this end, we conducted a theoretical analysis grounded in optimal transport theory (Santambrogio 2015) and demonstrated that the key bottleneck of ACKD lies in its

inherently higher cost of knowledge transfer. Compared to SCKD, the substantial semantic gap between input modalities leads to significantly increased transport costs during training. To further support this observation, we visualize the Wasserstein distance (Rubner, Tomasi, and Guibas 2000) in Figure 2, a widely used metric in optimal transport theory (Solomon et al. 2015; Chen et al. 2020; Frogner et al. 2015) that quantifies the cost of knowledge transfer across modalities. The results clearly show that ACKD incurs a much higher transport cost than SCKD. These findings highlight the urgent need for dedicated distillation frameworks tailored to ACKD, capable of bridging the semantic gap and enhancing cross-modal knowledge alignment.

To tackle the aforementioned challenges in ACKD, we propose SemBridge, a novel distillation framework designed to optimize knowledge transfer under semantic misalignment. Specifically, SemBridge integrates two plug-and-play modules: the Student-Friendly Matching (SFM) module and the Semantic-aware Knowledge Alignment (SKA) module. The SFM module aims to reduce transport costs by adaptively establishing suitable teacher-student matching. Inspired by the strong semantic correspondence typically assumed in SCKD, SFM first assigns an initial teacher to each student sample based on semantic similarity. Moreover, drawing inspiration from the human educational paradigm, SFM enables student samples to dynamically select their subsequent teachers throughout training based on evolving learning needs. In parallel, the SKA module is introduced to further optimize transport costs. It first formulates an intra-modal transport plan via Lagrangian optimization, capturing the semantic structure within each modality. Based on this, cross-modal transport plans are constructed separately for both the teacher and student modalities, facilitating more efficient and semantically aligned knowledge transfer.

Moreover, to facilitate our research, we construct a dataset benchmark with 3 sub-datasets, including MS images and asymmetric RGB images, namely S2S-EU, S2S-CN, and M2S-GL, respectively. The dataset includes a total of 70,414 MS images and 63,549 unpaired RGB images across diverse scene categories on Earth. To evaluate the generalization capability of SemBridge, we use MS images collected by different equipment with various numbers of spectral channels.

In experiments, we evaluate SemBridge under both homogeneous and heterogeneous model architectures by distilling knowledge from both multi-label and single-label teachers into single-label students. The results show that SemBridge yields consistent performance improvements across seven mainstream SCKD approaches in ACKD settings.

Our contribution can be summarized as: (1) To the best of our knowledge, we are the first to explore Asymmetric Cross-modal Knowledge Distillation (ACKD), a promising concept that broadens the application scope of Symmetric Cross-modal Knowledge Distillation (SCKD). (2) We propose SemBridge, a plug-and-play framework including Student-Friendly Matching (SFM) module and Semantic-aware Knowledge Alignment (SKA) module, to optimize the transport cost. (3) We construct a new benchmark with MS and asymmetric RGB image pairs to facilitate evaluation under real-world asymmetry.

Symbol	Description
$\mathcal{D}, \mathcal{D}_{match}$	Unpaired and Matched dataset
V, G, \tilde{G}	MS, RGB and Pseudo-RGB modality
T, S	Teacher and student
f_T, f_S	Feature extractors
h_T, h_S	Classifiers
M	Matcher
M_V, M_G	MS and RGB Encoder of M
z_T, z_S	Unfused features
\bar{z}_T, \bar{z}_S	Fused features
$\mathbf{p}_T, \mathbf{p}_S$	Outputs logits
v, \tilde{g}	Representation of MS and Pseudo-RGB
Planner	the proposed Planner

Table 1: Summary of Notations

Related Works

Remote Sensing (RS) Scene Classification aims to categorize geographic areas based on their semantic content. Early approaches relied on handcrafted features extracted from RGB images (Cheriyadat 2013; Zhang et al. 2013). Recently, deep learning methods have achieved notable success due to the strong generalization capability of neural networks (Zou et al. 2015; Cheng, Zhou, and Han 2016). However, in complex scenes, simply increasing network width or depth does not always improve performance, as RGB images often suffer from low information density. To address this, multispectral (MS) images have been introduced (Gómez and Meoni 2021), offering richer information via additional spectral bands. While MS images generally outperform RGB ones, their acquisition requires specialized sensors, and the increased spectral channels lead to higher computational costs. To alleviate these issues, researchers (Liu, Qu, and Zhang 2022; Shin et al. 2023) have proposed cross-modal knowledge distillation (CMKD) to transfer semantic knowledge from MS to RGB images, enabling efficient inference using only the RGB modality.

Symmetric modality-based Knowledge Distillation. Knowledge distillation (KD) was first proposed by Hinton *et al.* (Hinton, Vinyals, and Dean 2015) for optimizing the computational cost and memory consumptions on devices with limited computation or storage resources, which is regarded as uni-modality-based KD as both the teacher and student take the same modality as input. KD can be categorized into response-based (Sun et al. 2024; Zhao et al. 2022; Ba and Caruana 2014; Li et al. 2023; Hao et al. 2023; Hinton, Vinyals, and Dean 2015), feature-based (Yang et al. 2024b,a) as well as relation-based (Park et al. 2019; Yang et al. 2022, 2023) KD determined by which parts of the model are distilled. With the remarkable success of cross-modal learning (Kaur, Pannu, and Malhi 2021; He et al. 2016b), symmetric cross-modal knowledge distillation (SCKD) has gathered much attention, aiming to conduct knowledge from discriminate modalities to the weaker ones.

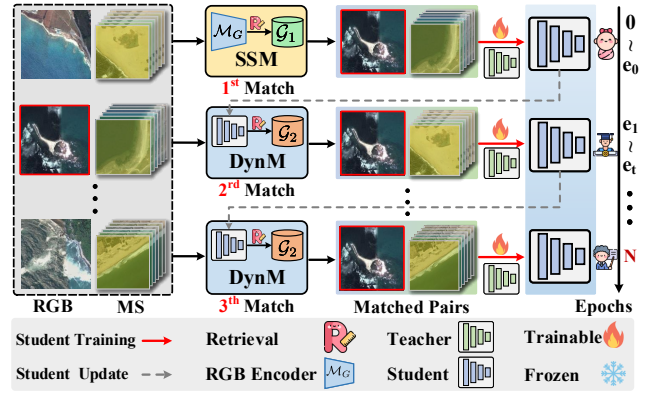


Figure 3: Illustration of the proposed student friendly matching strategy consisting of SSM for the first matching and DynM allowing student select proper teacher samples dynamically all the training stage. In DynM, current student model is designed to be involved.

Methodology

Overview

Given a teacher model T and a student model S taking MS modality V , and RGB modality G respectively. The dataset \mathcal{D} contains unpaired samples from C classes: $\mathcal{D} = \{(V_k^c)_{k=1}^{K_c}, (G_n^c)_{n=1}^{N_c}\}_{c=1}^C$, where c denotes the class index, and K_c, N_c are the number of MS and RGB samples in the c -th class, respectively. Let f_T and f_S denote the feature extractors of T and S , and h_T, h_S be their respective classifiers. Let $z_T = f_T(V)$, $z_S = f_S(G)$ be unfused features, and \bar{z}_T, \bar{z}_S be fused feature representations, obtained by applying adaptive average pooling on z_T and z_S . $\mathbf{p}_T = h_T(\bar{z}_T)$, $\mathbf{p}_S = h_S(\bar{z}_S)$ are output logits. We design a matcher $M = (M_V, M_G)$, which consisted of 2 encoders M_V and M_G to project V and pseudo-RGB images \tilde{G} to corresponding representations $v = M_V(V)$, $\tilde{g} = M_G(\tilde{G})$.

T is first trained in an offline manner. During this time, a matcher M is also trained for the initial matching, which will be introduced in the next sections.

At the training stage of S , we propose SemBridge consisting of a Student-Friendly Matching (SFM) module and a Semantic-aware Knowledge Alignment (SKA) module to select a teacher-student sample with greater semantic consistency and updated by the current student model several times. Finally, we finalize an optimal transport plan for weak semantic consistency modalities via the SKA module.

Optimal Transport Analysis

To demonstrate the knowledge transport cost caused by semantic consistency, we utilize the Wasserstein distance to compare the output logits extracted from strong (SCKD) and weak (ACKD) semantic consistency, respectively. Wasserstein distance is a common tool to measure optimal transport, which is used to evaluate the difficulty of knowledge propagation. Suppose x_S and x_T are the inputs of two modalities, the corresponding probability distributions:

Subset	S2S-EU	S2S-CN	M2S-GL
MS Label	Single	Single	Multiple
RGB Label	Single	Single	Single
Devices	Sentinel-2	Tiangong-2	Sentinel-2
MS bands	10	14	10
Resolution	64 × 64	128 × 128	120 × 120
Categories	10	10	15

Table 2: Details of the proposed dataset benchmark.

$z_s = f_S(x_S) \sim \mathcal{P}_S$ and $z_T = f_T(x_T) \sim \mathcal{P}_T$. The Wasserstein distance \mathbf{w} can be formulated as:

$$\mathbf{w}(\mathcal{P}_S, \mathcal{P}_T) = \inf_{\pi \in \Pi(\mathcal{P}_S, \mathcal{P}_T)} E_{(x_S, x_T) \sim \pi} [c(z_S, z_T)], \quad (1)$$

where $c(\cdot)$ is the distance measurement. z_S and z_T are the feature representations. $\Pi(\mathcal{P}_S, \mathcal{P}_T)$ is a set of candidate point of $\mathcal{P}_S(x_S)$ and $\mathcal{P}_T(x_T)$. $\Pi(\mathcal{P}_S, \mathcal{P}_T)$ is a joint distribution satisfying:

$$\begin{aligned} \int \pi(x_S, x_T) dx_T &= \mathcal{P}_S(x_S), \\ \int \pi(x_S, x_T) dx_S &= \mathcal{P}_T(x_T). \end{aligned} \quad (2)$$

As shown in Figure 2, due to significant cost, ACKD becomes more challenging compared to SCKD. Therefore, we are committed to finding a reasonable transport plan π to optimize the cost of knowledge propagation.

Student-Friendly Matching (SFM)

The first step to optimize the cost is to select a suitable teacher sample for each student with greater semantic consistency motivated by SCKD. Then, inspired by human educational wisdom, dynamic matching is proposed to select different teacher samples for students during their learning period, as shown in Figure 3. Specifically, by matching reasonable teacher samples for students, the cost can be optimized as:

$$\pi(i) = \arg \min_j \|z_S^i - z_T^j\|^2 \quad (3)$$

In other words, an optimal joint distribution is found by selecting teacher samples as Equ. (4):

$$\pi^*(x_S, x_T) = \begin{cases} 1, & \text{if } x_T = \arg \min_{x'_T} \|z_S - z'_T\|^2 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Subsequently, the optimized \mathbf{w} can be implemented as:

$$\mathbf{w}(\mathcal{P}_S, \mathcal{P}_T) = \sum_i \pi^*(x_S^i, x_T^j) \|z_S^i - z_T^j\|^2. \quad (5)$$

Construct Retrieval Teacher Galleries. Before training S , we construct two teacher galleries, noted as \mathcal{G}_1 and \mathcal{G}_2 , to enable the student to retrieve teacher samples in SFM. The trained MS encoder M_V and T are used to project MS samples into embeddings and logits and saved in \mathcal{G}_1 and \mathcal{G}_2 respectively. Specifically, for the c -th class, let: $V^c = \{V_k^c\}_{k=1}^{K_c}$ denote the MS samples. M_V and T project it into: $v_k^c = M_V(V_k^c)$, $\mathbf{p}_{T,k}^c = T(V_k^c)$, $\forall k = 1, \dots, K_c$, (6) and save v_k^c and $\mathbf{p}_{T,k}^c$ in \mathcal{G}_1 and \mathcal{G}_2 respectively. The saved galleries will be utilized in the subsequent matching stages.

Self-supervised Semantic-aware Matching (SSM). To learn a semantic-aware matcher without relying on paired RGB images, we utilize only V from the unpaired dataset \mathcal{D} . Specifically, we extract the RGB bands from each V to construct a pseudo-RGB image \tilde{G} . Since V and \tilde{G} originate from the same source, they naturally share the same semantic content and are treated as positive pairs for self-supervised learning (Jing and Tian 2020). It should be noted that the split channels are determined by the modality of the student model. For example, in this task, the student modality is the RGB image. Hence, we just split R, G, and B channels from V .

Then, InfoNCE loss (Gutmann and Hyvärinen 2010) is employed to optimize the matcher to learn the semantic difference in Contrastive Language–Image Pretraining (CLIP)-based manner (Radford et al. 2021).

After training M , we use it to select the most semantically consistent teacher samples for each student sample within the same class. For the c -th class, let: $G^c = \{G_n^c\}_{n=1}^{N_c}$ denote RGB samples and projected into $g_n^c = M_G(G_n^c)$. Then, for each g_n^c , we compute its cosine similarity $\mathbf{cos}(\cdot, \cdot)$ with all teacher embeddings v_k^c from \mathcal{G}_1 to form a similarity matrix:

$$\Phi_n^c = [\mathbf{cos}(g_n^c, v_1^c), \mathbf{cos}(g_n^c, v_2^c), \dots, \mathbf{cos}(g_n^c, v_{K_c}^c)]. \quad (7)$$

By stacking all similarity vectors, we obtain the class-wise similarity matrix:

$$\Phi^c = [\Phi_1^c, \Phi_2^c, \dots, \Phi_{N_c}^c]^T \in R^{N_c \times K_c}. \quad (8)$$

Next, for each student sample G_n^c with embedding g_n^c , we select the teacher sample with the highest semantic similarity: $\mathbf{k}^* = \arg \max_{\mathbf{k}} \Phi_{n,\mathbf{k}}^c$. This yields the matched sample pairs for the c -th class:

$$\mathcal{D}_{\text{match}} = \{(V_{\mathbf{k}^*}^c, G_n^c) \mid n = 1, \dots, N_c\}. \quad (9)$$

Dynamic Matching (DyM). Inspired by human education systems where students are guided by different teachers throughout their learning journey, we propose a DyM strategy. Instead of relying on a fixed teacher, DyM periodically updates the matched teacher-student pairs during training. This allows the student to absorb knowledge from multiple teacher samples, thereby reducing semantic bias and improving generalization. First, we compute the output logits of the student $\mathbf{p}_{S,n}^c = S(G_n^c)$ from the c -th class. Then, we calculate the Kullback-Leibler (KL) divergence with temperature γ between the n^{th} student prediction $\mathbf{p}_{S,n}^c$ and all candidate teacher samples $\mathbf{p}_{T,k}^c$ from \mathcal{G}_2 :

$$\Omega_n^c = \{\text{KL}(\mathbf{p}_{S,n}^c \parallel \mathbf{p}_{T,k}^c; \gamma)\}_{k=1}^{K_c}. \quad (10)$$

Unlike selecting teacher samples with maximum semantic similarity to acquire basic knowledge in the early stage of learning at SSM, DyM encourages the student to select more challenging samples, facilitating a progressive transition from easy to difficult knowledge:

$$\mathbf{k}^* = \arg \min_{\mathbf{k}} \Omega_{n,\mathbf{k}}^c. \quad (11)$$

The time of selecting new teachers in human educational systems almost depends on the years of study in the current

	Model	S2S-EU		S2S-CN		M2S-GL	
		OA	F1	OA	F1	OA	F1
Homogeneous model	T:ResNet34	95.3	95.1	96.8	97.0	/	80.8
	S:ResNet34	91.7	91.6	94.9	94.4	94.9	93.2
	+SemBridge	93.7	93.6	96.2	95.8	96.6	95.1
	T:MobileNetV2	95.2	95.0	95.3	95.5	/	75.2
	S:MobileNetV2	89.4	89.1	92.3	91.3	92.9	90.3
	+SemBridge	91.7	91.5	93.6	92.8	93.9	91.7
Homogeneous model	T:ShuffleNetV2	92.3	92.0	93.7	93.5	/	70.3
	S:ShuffleNetV2	85.9	85.6	90.0	88.8	88.8	85.5
	+SemBridge	88.4	88.1	91.4	90.6	90.8	87.8
	T:ResNet34	95.3	95.1	96.8	97.0	/	80.8
	S:MobileNet	89.4	89.1	92.3	91.3	92.9	90.3
	+SemBridge	92.1	91.9	93.5	92.9	93.9	91.7
Heterogeneous model	T:ResNet34	95.3	95.3	96.8	96.8	/	85.3
	S:ShuffleNetV2	85.9	85.6	90.0	88.8	88.8	85.5
	+SemBridge	87.9	87.6	91.0	89.8	90.3	87.8
	T:MobileNetV2	95.2	95.0	95.3	95.5	/	75.2
	S:ShuffleNetV2	85.9	85.6	90.0	88.8	88.8	85.5
	+SemBridge	87.8	87.4	91.6	90.7	89.7	87.5

Table 3: Compared to the Baseline without KD. ‘T’ and ‘S’ denote the teacher and student model, respectively.

stage. Based on this, DynM is performed several times along the learning journey as shown in Figure 3. Inspired by curriculum learning (Bengio et al. 2009), the time of per matching is gradually extended with the increment of knowledge diversity and implemented as:

$$e_t = e_0 + \sum_{i=1}^t (\Delta e + e_\mu(i-1)). \quad (12)$$

Here, t is the number of times to perform DynM and e_0 is the initial matching time (epoch).

Semantic-aware Knowledge Alignment (SKA)

To optimize the transport cost between matched samples, in this section, a transport plan π is finalized, so we name this module as Planner as shown in Figure 4. Suppose the overall transport cost of two distributions x and y containing m and n samples respectively:

$$\mathcal{L}_{OT} = \sum_{i=1}^m \sum_{j=1}^n \pi_{ij} c(x_i, y_j). \quad (13)$$

To estimate the optimal transport plan $\pi_{x \rightarrow y}$ between x and y , we utilize Lagrangian functions with boundary regularization and entropy regularization. The details of this part can be found in Appendix A. Finally, intra-modality transport plan can be formulated as:

$$\pi_{x \rightarrow y} = \text{softmax}\left(\frac{c(x, y)}{\epsilon}\right). \quad (14)$$

To formulate $c(\cdot)$ and coefficient ϵ w.r.t. Equ. (14), we instantiate Equ. (14) with a learnable multi-head attention

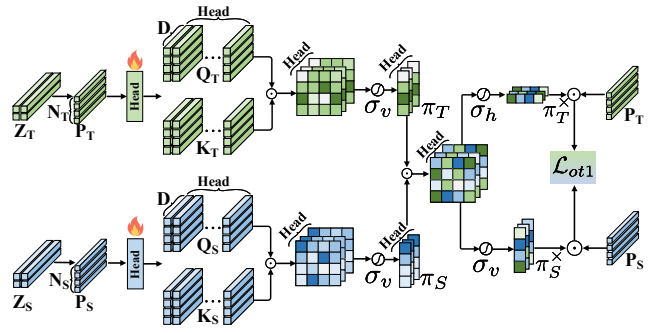


Figure 4: The structure of Planner.

structure to avoid manual choices of $c(\cdot)$ and ϵ , inspired by its resemblance to the formulation of multi-head attention noted as Planner as shown in Figure 4. Specifically, z_T and z_S are flattened with N patches as P_T and P_S before feeding into Planner. Planner project P_T and P_S into Q_T, K_T and Q_S, K_S with H heads as following:

$$Q_m, K_m = \text{Planner}(P_m), \quad m \in \{T, S\}, \quad (15)$$

and obtain their transport plan as $\pi_T, \pi_S \in R^{H \times N}$:

$$\pi_m = \text{softmax}\left(Q_m \cdot \frac{K_m^\top}{\sqrt{d}}\right), \quad m \in \{T, S\}, \quad (16)$$

where d is the feature dimension per head. Subsequently, cross-modality transmission plans are implemented as:

$$\begin{aligned} \pi_T^\times &= \frac{\sigma_h(\sigma_v(\pi_T) \cdot \sigma_v(\pi_S))}{\epsilon_T}, \\ \pi_S^\times &= \frac{\sigma_v(\sigma_v(\pi_S) \cdot \sigma_v(\pi_T))}{\epsilon_S}, \end{aligned} \quad (17)$$

where $\epsilon_T = \frac{1}{N_T} \sum_{i=1}^{N_T} P_T$ and $\epsilon_S = \frac{1}{N_S} \sum_{i=1}^{N_S} P_S$. ϵ_T and ϵ_S are scaling factors for numerical stability, computed from the average of patches P_T and P_S . N_T and N_S are the number of patches of the teacher and student samples, respectively, while σ_h and σ_v denote the horizontal and vertical mean pooling operations. The cross-modal transmission plan is then applied to z_T and z_S as follows:

$$D_T = z_T \cdot \frac{1}{H} \sum_{h=1}^H \pi_T^{\times, h}, \quad D_S = z_S \cdot \frac{1}{H} \sum_{h=1}^H \pi_S^{\times, h}. \quad (18)$$

Finally, to further bridge the modality gap, we employ CORAL, noted as Cor (Sun and Saenko 2016) to align refined feature D_T and D_S and fused feature \bar{z}_T and \bar{z}_S respectively and get cost \mathcal{L}_{ot1} and \mathcal{L}_{ot2} implemented as:

$$\mathcal{L}_{ot1} = \text{Cor}(D_T, D_S), \quad \mathcal{L}_{ot2} = \text{Cor}(\bar{z}_T, \bar{z}_S). \quad (19)$$

The overall loss function is formulated as:

$$\mathcal{L}_{all} = \mathcal{L}_{task} + \lambda_1 \mathcal{L}_{kd} + \lambda_2 (\mathcal{L}_{ot1} + \mathcal{L}_{ot2}). \quad (20)$$

Here, $\mathcal{L}_{kd} \in \{\text{Vanilla KD (Hinton, Vinyals, and Dean 2015), RKD (Park et al. 2019), DKD (Zhao et al. 2022), Logits (Ba and Caruana 2014), CTKD (Li et al. 2023), STKD (Sun et al. 2024)}\}$ denotes the SCKD loss and \mathcal{L}_{task} is the task-related loss. λ_1 and λ_2 are balanced factors. We set $\lambda_2 = 1 - \lambda_1$. The detail of CORAL is implemented in Appendix A.

Datasets	S2S-EU							S2S-CN							M2S-GL						
Methods	R/R	M/M	S/S	R/M	R/S	M/S	Avg.	R/R	M/M	S/S	R/M	R/S	M/S	Avg.	R/R	M/M	S/S	R/M	R/S	M/S	Avg.
RKD	91.6	89.1	85.9	89.9	85.1	86.0	87.9	94.9	91.9	90.0	91.7	90.0	89.7	91.4	95.1	<u>93.4</u>	88.3	<u>93.6</u>	88.9	<u>89.1</u>	91.4
DKD	91.7	91.7	<u>87.1</u>	90.4	86.2	<u>86.8</u>	89.0	94.9	93.1	89.8	92.2	89.5	90.7	91.7	68.1	63.6	64.9	63.5	59.1	60.0	63.2
Logits	87.8	89.1	73.8	88.8	84.0	85.9	84.9	94.7	92.6	90.0	91.6	90.5	90.7	91.7	93.6	91.2	85.0	91.4	86.1	88.1	89.2
CTKD	92.5	90.9	71.8	92.1	86.6	86.0	86.7	94.8	<u>93.5</u>	90.4	92.2	90.1	91.1	<u>92.0</u>	89.0	87.1	82.1	89.0	82.3	82.0	85.3
LSKD	92.1	88.8	86.8	89.9	85.5	85.8	88.2	95.4	91.9	89.6	91.4	91.0	90.8	91.7	<u>95.4</u>	93.3	<u>90.2</u>	<u>93.1</u>	<u>89.3</u>	<u>89.1</u>	<u>91.7</u>
VPR	46.2	75.2	33.9	69.2	38.8	40.1	50.6	94.4	90.0	88.5	90.1	88.4	88.7	90.0	94.1	91.2	86.7	91.0	87.9	85.7	89.4
\mathcal{L}_{kd}	<u>92.6</u>	89.3	86.2	90.1	85.3	85.3	88.1	<u>95.6</u>	91.9	89.8	<u>92.3</u>	89.8	<u>91.2</u>	91.8	93.6	91.5	84.6	90.8	85.5	84.5	88.4
+Ours	93.7	91.7	88.4	92.1	87.9	87.8	90.3	96.2	93.6	91.4	93.5	91.0	91.6	92.9	96.6	93.9	90.8	93.9	90.3	89.7	92.5

Table 4: Compared with SOTA methods in terms of OA. **R**, **M** and **S** indicates ResNet34, MobileNetV2 and ShuffleNetV2 respectively. \mathcal{L}_{kd} is based on Vanilla KD. The best results are marked in **bold** and the second best in underline.

Method	S2S-EU		S2S-CN		M2S-GL	
	OA	F1	OA	F1	OA	F1
Vanilla KD	92.6	92.3	95.6	95.0	93.6	91.6
+ SemBridge	93.7	93.6	96.2	95.8	96.6	95.1
RKD	91.6	91.5	94.9	94.3	95.1	93.2
+ SemBridge	92.3	92.2	95.7	95.3	95.4	93.6
DKD	91.7	91.5	94.9	94.3	68.1	73.5
+ SemBridge	92.4	92.2	95.3	95.0	83.0	82.9
Logits	87.8	87.5	94.7	94.2	93.6	91.2
+ SemBridge	91.4	91.2	96.0	95.5	94.3	92.4
CTKD	92.5	92.3	94.8	94.3	89.0	88.0
+ SemBridge	93.3	93.2	95.8	95.5	95.4	93.1
LSKD	92.1	92.0	95.4	95.0	95.4	93.1
+ SemBridge	92.7	92.5	95.9	95.5	95.5	93.4

Table 5: Generalization capability testing.

Dataset Construction

Lacking modalities with weak semantic consistency in RS scene classification tasks hampers the application of knowledge propagation. Therefore, a comprehensive modality paired with asymmetric information is indispensable. To address this issue, we construct a new dataset benchmark consisting of 3 sub-datasets, S2S-EU, S2S-CN, and M2S-GL with MS images and RGB images as shown in Table 2. Due to unique geographical environments, scenes of the same category in RS images always present various semantic content worldwide. The goal of this research is to propagate knowledge from any place or country to others, regardless of the semantic content. To do this, we investigated and collected available MS images from 3 public datasets, i.e., EuroSAT (Helber et al. 2018), NaSC-TG2 (Zhou et al. 2021), and BigEarthNet (Sumbul et al. 2019), respectively, which contain scenes from around the world. Subsequently, we collected RGB images from other public datasets as an asymmetric modality. The details of the proposed dataset bench-

mark can be found in Appendix B. Finally, to evaluate the difficulty of knowledge distillation in ACKD, we compute the mutual information (Batina et al. 2011) within class on 3 proposed datasets, which is shown in Figure 6 in Appendix.

Experiments

Experimental Setup

Datasets. Self-constructed benchmarks involving S2S-EU, S2S-CN, and M2S-GL are employed to evaluate SemBridge for RS scene classification tasks. Specifically, S2S-EU and S2S-CN are used to evaluate the performance in single-label→single-label classification, while M2S-GL is employed to assess knowledge propagation from multi-label→single-label classification.

Evaluation metrics. Overall Accuracy (OA) and F1-score (F1) are utilized to evaluate classification performance. Following the setup in (Liu, Qu, and Zhang 2022), only F1 is used to evaluate teacher performance on multi-label classification tasks in M2S-GL.

Compared method. We compared several methods in this section. ‘Baseline’ denotes the original training without KD. We also employ SCKD approaches Vanilla KD (Hinton, Vinyals, and Dean 2015), RKD (Park et al. 2019), DKD (Zhao et al. 2022), Logits (Ba and Caruana 2014), CTKD (Li et al. 2023), LSKD (Sun et al. 2024)}, and VPR (Wang et al. 2024) to evaluate the performance on ACKD compared with applying the proposed SemBridge(+SemBridge).

Evaluated Network. Experiment are conducted over ResNet34 (He et al. 2016a), MobileNetV2 (Sandler et al. 2018) and ShuffleNetV2 (Ma et al. 2018). The whole training details is implemented in Appendix C.

Compared with Baseline Methods

In Table 3, we conduct experiments on both homogeneous and heterogeneous model architectures. Compared to baseline without KD, the SemBridge with Vanilla KD enables the student model to achieve significant improvements across 3 datasets. For example, for a homogeneous model of ResNet34, SemBridge leads to 1.3%~2.0% and

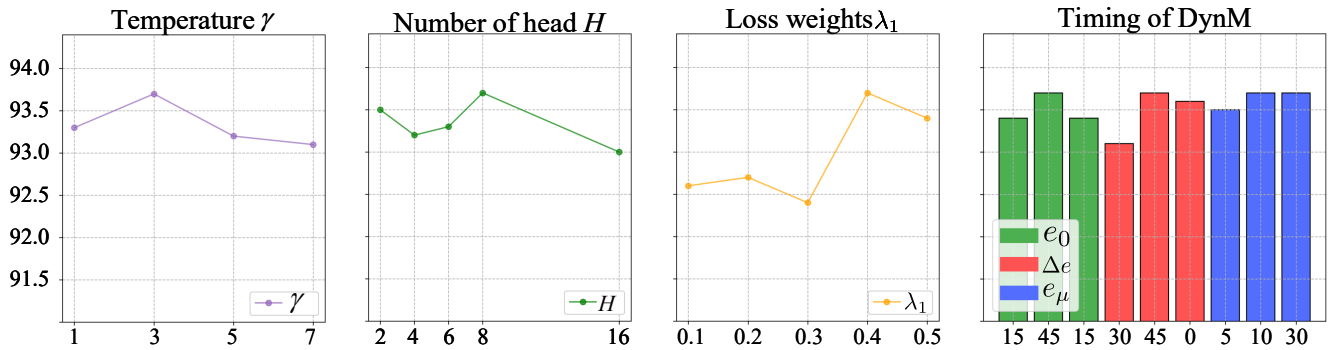


Figure 5: Hyperparameter analysis.

1.4%~2.0% gains on S2S-EU, S2S-CN, and M2S-GL in terms of OA and F1, respectively. Furthermore, to evaluate the performance of SemBridge under the different architectures between the teacher and the student, ShuffleNetV2 and MobileNetV2 are supervised by homogeneous and heterogeneous teachers, respectively. The results indicate that knowledge can be propagated effectively via SemBridge regardless of model architectures.

Compared with State-of-the-art Methods

Table 4 reports the classification performance of SemBridge based on Vanilla KD across 6 combinations of model architecture. SemBridge enables Vanilla KD to achieve SOTA performance on 3 datasets in terms of OA. For ResNet34, SemBridge enables Vanilla KD to achieve improvements of 0.6% and 1.1% on S2S-EU and S2S-CN, and outperforms LSKD by 1.2% on M2S-GL. For ShuffleNetV2 supervised by ResNet34, our approach outperforms CTKD and LSKD with gains of 1.3% and 1.0% on S2S-EU and M2S-GL, respectively. Compared to uni-modality-based methods (Vanilla KD, RKD, DKD, Logits, LSKD, CTKD), VPR is designed to distill knowledge between modalities with the same semantic content. It can be found that due to semantic differences, VPR shows unpromising results, especially on S2S-EU. It also indicates the necessity of ACKD.

Generalization Capability Testing

Table 5 illustrates the generalization capability testing on ResNet34. It can be observed that SemBridge can enhance the performance of existing SCKD approaches on ACKD tasks. Compared to others, SemBridge with Vanilla KD achieves the best performance with 93.7%~96.6% and 93.6%~95.8% in terms of OA and F1, respectively. On S2S-EU, SemBridge shows the greatest improvement based on Logits with gains of 3.6% and 3.7% in terms of OA and F1, respectively. On multi-label→single-label tasks, SemBridge achieves the largest improvement for DKD, with increases of 14.9% and 9.4% of OA and F1, respectively. Please note that we only applied SemBridge to uni-modality-based methods, which are often used on more universal scenarios.

SSM	DynM	\mathcal{L}_{ot1}	\mathcal{L}_{ot2}	S2S-EU	S2S-CN	M2S-GL
✗	✓	✓	✓	92.5	95.3	95.6
✓	✗	✓	✓	92.9	95.1	94.2
✓	✓	✗	✓	92.5	96.1	95.1
✓	✓	✓	✗	92.8	94.1	95.8
✓	✓	✓	✓	93.7	96.2	96.6

Table 6: Impact of SSM, DynM, and SKA(\mathcal{L}_{ot1} , \mathcal{L}_{ot2}) of SemBridge on R/R in terms of OA.

Hyperparameter Analysis and Ablation Study

As shown in Figure 5, we analyze the effects of temperature γ in the matcher, the number of heads H in Planner, and the loss weight λ_1 . The optimal performance is observed when $\gamma = 3$, $H = 8$ and $\lambda_1 = 0.4$. Furthermore, the timing of DynM, i.e., e_0 , Δe , and e_μ exhibit consistent robustness across settings. Besides, we investigate the effectiveness of each component in SemBridge as shown in Table 6. All four components contribute to the best result, which is 93.7% on S2S-EU, 96.2% on S2S-CN, and 96.6% on M2S-GL, indicating that SemBridge can optimize the cost caused by weak semantic consistency. The impact of DynM and the Planner on optimal transport can be found in Appendix C.

Conclusion

In this paper, we propose ACKD, a new research direction to broaden the application scope of SCKD. To this end, we construct a dataset benchmark comprising 3 sub-datasets in the remote sensing fields. Subsequently, we propose a framework, namely SemBridge, consisting of a Student-Friendly Matching module and a Semantic-Aware Knowledge Alignment module to reduce the transport cost during knowledge distillation. The experimental results demonstrate that the proposed SemBridge not only helps Vanilla KD achieve state-of-the-art performance across various datasets but also enhances the performance of existing SCKD methods on ACKD, indicating superior generalization capability. However, we also identify some limitations of SemBridge. The time consumption associated with student-friendly matching may negatively impact training speed. We regard it as the future direction.

Acknowledgments

This work is supported by the National Key R&D Program of China (2022YFB4501600) and the National Natural Science Foundation of China (62406312).

References

- Ba, J.; and Caruana, R. 2014. Do deep nets really need to be deep? *Advances in neural information processing systems*, 27.
- Batina, L.; Gierlichs, B.; Prouff, E.; Rivain, M.; Standaert, F.-X.; and Veyrat-Charvillat, N. 2011. Mutual information analysis: a comprehensive study. *Journal of Cryptology*, 24(2): 269–291.
- Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.
- Chen, L.; Gan, Z.; Cheng, Y.; Li, L.; Carin, L.; and Liu, J. 2020. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, 1542–1553. PMLR.
- Cheng, G.; Han, J.; and Lu, X. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10): 1865–1883.
- Cheng, G.; Zhou, P.; and Han, J. 2016. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE transactions on geoscience and remote sensing*, 54(12): 7405–7415.
- Cheriyadat, A. M. 2013. Unsupervised feature learning for aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1): 439–451.
- Dai, R.; Das, S.; and Bremond, F. 2021. Learning an augmented rgb representation with cross-modal knowledge distillation for action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13053–13064.
- Frogner, C.; Zhang, C.; Mobahi, H.; Araya, M.; and Poggio, T. A. 2015. Learning with a Wasserstein loss. *Advances in neural information processing systems*, 28.
- Gómez, P.; and Meoni, G. 2021. MSMatch: Semisupervised multispectral scene classification with few labels. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 11643–11654.
- 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.
- Hao, Z.; Guo, J.; Han, K.; Tang, Y.; Hu, H.; Wang, Y.; and Xu, C. 2023. One-for-All: Bridge the Gap Between Heterogeneous Architectures in Knowledge Distillation. In *Advances in neural information processing systems*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- He, Y.; Xiang, S.; Kang, C.; Wang, J.; and Pan, C. 2016b. Cross-modal retrieval via deep and bidirectional representation learning. *IEEE Transactions on Multimedia*, 18(7): 1363–1377.
- Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2018. Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS 2018-2018 IEEE international geoscience and remote sensing symposium*, 204–207. IEEE.
- Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv:1503.02531*.
- Huo, F.; Xu, W.; Guo, J.; Wang, H.; and Guo, S. 2024. C2KD: Bridging the Modality Gap for Cross-Modal Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16006–16015.
- Jing, L.; and Tian, Y. 2020. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11): 4037–4058.
- Kaur, P.; Pannu, H. S.; and Malhi, A. K. 2021. Comparative analysis on cross-modal information retrieval: A review. *Computer Science Review*, 39: 100336.
- Kettig, R. L.; and Landgrebe, D. 1976. Classification of multispectral image data by extraction and classification of homogeneous objects. *IEEE Transactions on Geoscience Electronics*, 14(1): 19–26.
- Kim, S.; Kim, Y.; Hwang, S.; Jeong, H.; and Kum, D. 2024. LabelDistill: Label-Guided Cross-Modal Knowledge Distillation for Camera-Based 3D Object Detection. In *European Conference on Computer Vision*, 19–37. Springer.
- Landgrebe, D. 2002. Hyperspectral image data analysis. *IEEE Signal processing magazine*, 19(1): 17–28.
- Li, M.; Zhang, Y.; Xie, Y.; Gao, Z.; Li, C.; Zhang, Z.; and Qu, Y. 2022. Cross-domain and cross-modal knowledge distillation in domain adaptation for 3d semantic segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3829–3837.
- Li, Z.; Li, X.; Yang, L.; Zhao, B.; Song, R.; Luo, L.; Li, J.; and Yang, J. 2023. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1504–1512.
- Liu, H.; Qu, Y.; and Zhang, L. 2022. Multispectral Scene Classification via Cross-Modal Knowledge Distillation. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–12.
- Lu, A.; Zhao, J.; Li, C.; Xiao, Y.; and Luo, B. 2024. Breaking Modality Gap in RGBT Tracking: Coupled Knowledge Distillation. In *ACM Multimedia 2024*.
- Ma, F.; Yuan, M.; and Kozak, I. 2023. Multispectral imaging: Review of current applications. *Survey of Ophthalmology*, 68(5): 889–904.
- Ma, N.; Zhang, X.; Zheng, H.-T.; and Sun, J. 2018. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, 116–131.

- Park, J.-I.; Lee, M.-H.; Grossberg, M. D.; and Nayar, S. K. 2007. Multispectral imaging using multiplexed illumination. In *2007 IEEE 11th International Conference on Computer Vision*, 1–8. IEEE.
- Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3967–3976.
- Phiri, D.; and Morgenroth, J. 2017. Developments in Landsat land cover classification methods: A review. *Remote Sensing*, 9(9): 967.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.
- Ren, S.; Du, Y.; Lv, J.; Han, G.; and He, S. 2021. Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13325–13333.
- Reutebuch, S. E.; Andersen, H.-E.; and McGaughey, R. J. 2005. Light detection and ranging (LIDAR): an emerging tool for multiple resource inventory. *Journal of forestry*, 103(6): 286–292.
- Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover’s distance as a metric for image retrieval. *International journal of computer vision*, 40: 99–121.
- Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.
- Santambrogio, F. 2015. *Optimal transport for applied mathematicians*, volume 87. Springer.
- Sarkar, P.; and Etemad, A. 2024. Xkd: Cross-modal knowledge distillation with domain alignment for video representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14875–14885.
- Shin, H.-K.; Uhm, K.-H.; Jung, S.-W.; and Ko, S.-J. 2023. Multispectral-to-RGB Knowledge Distillation for Remote Sensing Image Scene Classification. *IEEE Geoscience and Remote Sensing Letters*, 20: 1–5.
- Solomon, J.; De Goes, F.; Peyré, G.; Cuturi, M.; Butscher, A.; Nguyen, A.; Du, T.; and Guibas, L. 2015. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (ToG)*, 34(4): 1–11.
- Sumbul, G.; Charfuelan, M.; Demir, B.; and Markl, V. 2019. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, 5901–5904. IEEE.
- Sun, B.; and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, 443–450. Springer.
- Sun, S.; Ren, W.; Li, J.; Wang, R.; and Cao, X. 2024. Logit standardization in knowledge distillation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 15731–15740.
- Wang, S.; She, R.; Kang, Q.; Jian, X.; Zhao, K.; Song, Y.; and Tay, W. P. 2024. DistilVPR: Cross-Modal Knowledge Distillation for Visual Place Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 10377–10385.
- Wang, W.; Liu, F.; Liao, W.; and Xiao, L. 2023. Cross-modal graph knowledge representation and distillation learning for land cover classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–18.
- Xue, Z.; Gao, Z.; Ren, S.; and Zhao, H. 2022. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. *arXiv preprint arXiv:2206.06487*.
- Yang, C.; An, Z.; Huang, L.; Bi, J.; Yu, X.; Yang, H.; Diao, B.; and Xu, Y. 2024a. CLIP-KD: An Empirical Study of CLIP Model Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15952–15962.
- Yang, C.; An, Z.; Zhou, H.; Zhuang, F.; Xu, Y.; and Zhang, Q. 2023. Online knowledge distillation via mutual contrastive learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 10212–10227.
- 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.
- Yang, Z.; Li, Z.; Zeng, A.; Li, Z.; Yuan, C.; and Li, Y. 2024b. ViTKD: Feature-based Knowledge Distillation for Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1379–1388.
- Zhang, Y.; Sun, X.; Wang, H.; and Fu, K. 2013. High-Resolution Remote-Sensing Image Classification via an Approximate Earth Mover’s Distance-Based Bag-of-Features Model. *IEEE Geoscience and Remote Sensing Letters*, 10(5): 1055–1059.
- Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled knowledge distillation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11953–11962.
- Zhao, H.; Zhang, Q.; Zhao, S.; Chen, Z.; Zhang, J.; and Tao, D. 2024. SimDistill: Simulated Multi-Modal Distillation for BEV 3D Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7460–7468.
- Zhou, Z.; Li, S.; Wu, W.; Guo, W.; Li, X.; Xia, G.; and Zhao, Z. 2021. NaSC-TG2: Natural Scene Classification With Tiangong-2 Remotely Sensed Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 3228–3242.
- Zou, Q.; Ni, L.; Zhang, T.; and Wang, Q. 2015. Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and remote sensing letters*, 12(11): 2321–2325.