CAUSAL KNOWLEDGE DISTILLATION FOR ROBUST REMOTE SENSING SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation (KD) is widely used to compress large remote sensing semantic segmentation models for deployment on resource-constrained satellites. However, variations across satellite sensors and environmental conditions in real-world deployment can induce significant domain shifts between the source domain and unseen deployment domains. Existing KD methods primarily optimize source-domain accuracy while often overlooking robustness under domain shifts. To improve domain generalization in remote sensing semantic segmentation, we revisit KD from a causal perspective and attribute its limited robustness to two sources: domain-dependent noncausal knowledge transferred by the teacher, as well as noncausal shortcuts acquired by the student from source-domain supervision. Specifically, we propose Causal Knowledge Distillation (CKD), comprising Teacher Causal Knowledge Transfer (TCKT) and Student Causal Knowledge Acquisition (SCKA). TCKT learns prompts to refine teacher representations by suppressing domain-sensitive responses, thereby reducing the transfer of noncausal knowledge during distillation. SCKA applies causal frequency intervention to student inputs to mitigate the acquisition of domain-related shortcuts and encourage stable causal representation learning. In addition, CKD can be readily integrated with different distillation objectives without modifying the student architecture. Extensive experiments across multiple domain shifts demonstrate that integrating CKD with existing KD methods consistently improves robustness.

1 INTRODUCTION

Remote sensing semantic segmentation is widely used in applications such as land-cover mapping, urban planning, and disaster assessment (Demir et al., 2018; Robinson et al., 2019; Wang et al., 2021). However, deploying segmentation models on satellites requires lightweight architectures due to limited onboard computation, memory, and power (Bahl et al., 2019). Knowledge distillation

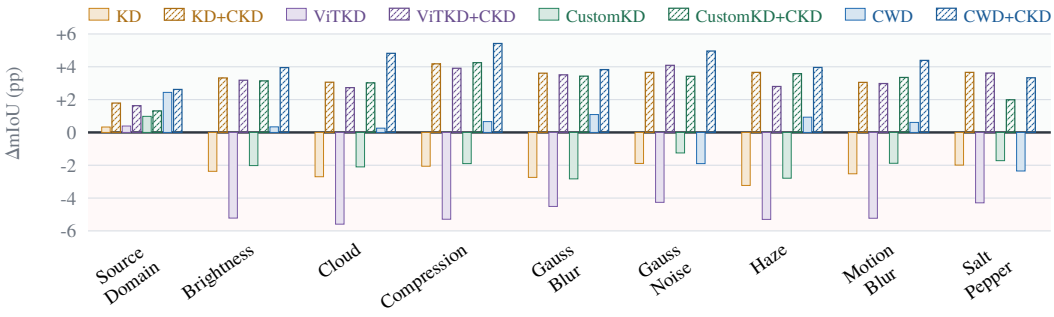


Figure 1: Comparison of conventional KD methods and their CKD-augmented variants on the source-domain test set of the ISPRS Potsdam dataset (Rottensteiner et al., 2012) and eight domain-shift test domains from REOBench (Li et al., 2025). For each domain, values report the mIoU difference between the corresponding distilled model and the undistilled student. The results show that conventional KD improves source-domain performance but is less robust under domain shifts, whereas CKD-augmented variants improve robustness without sacrificing source-domain accuracy.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

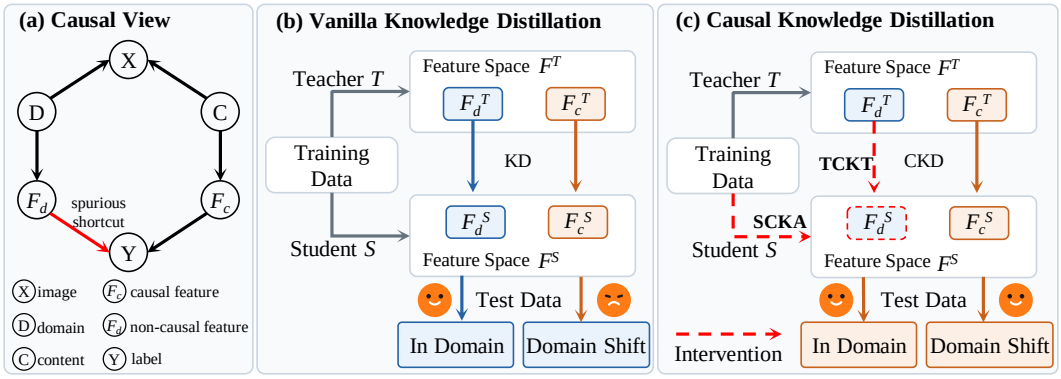


Figure 2: Motivation of our CKD. (a) Content C induces causal semantic features F_c that remain stable across domains, whereas domain factors D induce domain-dependent noncausal features F_d. (b) Vanilla KD transfers both causal features F_c^T and domain-dependent features F_d^T from the teacher, while source-domain supervision can independently induce noncausal features F_d^S in the student, resulting in limited robustness to domain shifts. (c) CKD addresses these two sources of noncausal bias: TCKT suppresses the transfer of F_d^T, while SCKA reduces the acquisition of F_d^S, encouraging the student to rely on causal semantic features and improving robustness under domain shifts.

(KD) addresses this challenge by transferring knowledge from a high-capacity teacher to a compact student. However, in real-world deployment, variations across satellite sensors, atmospheric conditions, and imaging environments are common. These factors lead to significant distribution shifts between the source domain used for training and unseen deployment domains.

Existing KD methods (Liang et al., 2024; Sun et al., 2025) primarily optimize the transfer of source-domain knowledge, with limited consideration of robustness to domain shifts commonly encountered in remote sensing, as illustrated in Figure 1. To improve generalization, recent studies transfer general-purpose representations from vision foundation model (VFM) teachers to compact students before downstream adaptation (Huang et al., 2023; Lv et al., 2026). However, domain-specific bias can still enter the student through two complementary routes. First, a VFM teacher encodes domain-dependent factors (Zhang et al., 2026) and transfers them together with task-relevant knowledge during distillation. Second, the student acquires spurious cues from source-domain remote sensing data. Therefore, robust knowledge distillation for remote sensing must address both the *transfer* and *acquisition* of noncausal knowledge.

Motivated by these observations, we revisit robust distillation from a causal perspective, as illustrated in Figure 2. We argue that predictions should rely on causal semantic factors that remain stable across domains, rather than domain-dependent noncausal factors. Accordingly, we address two complementary problems: transferring causal knowledge from the teacher and inducing the student to learn causal knowledge from source data. To this end, we propose Causal Knowledge Distillation (CKD), comprising Teacher Causal Knowledge Transfer (TCKT) and Student Causal Knowledge Acquisition (SCKA). TCKT uses learnable frequency prompts to refine teacher representations and suppress domain-sensitive knowledge before distillation, whereas SCKA applies frequency interventions that preserve causal semantic information while reducing the student's reliance on domain-specific cues during learning. In short, TCKT controls *what* is transferred, whereas SCKA controls *how* the student learns. More importantly, CKD can be integrated into existing KD objectives without modifying the student architecture or introducing additional inference cost.

Our contributions are threefold. (1) We formulate robust knowledge distillation from a causal perspective and identify two sources of non-causal knowledge: that transferred by the teacher and that independently acquired by the student. (2) We propose TCKT and SCKA to address these two sources of noncausal bias. TCKT refines teacher features to promote causal knowledge transfer, while SCKA applies student frequency intervention to encourage causal knowledge acquisition. (3) Both modules are architecture agnostic, compatible with diverse distillation objectives, and introduce no additional inference cost. Experiments across multiple domain-shift test sets demonstrate improvements in robustness under domain shift.

2 Related Work

Remote Sensing Segmentation under Domain Shift. Remote sensing semantic segmentation is highly vulnerable to domain shifts caused by variations in geographic regions, imaging platforms, sensors, weather, and acquisition quality. Existing benchmarks have revealed substantial geographic and sensor-induced domain gaps, motivating research on cross-domain remote sensing segmentation (Wang et al., 2021; Shermeyer et al., 2020). Unlike domain adaptation, which assumes access to a predefined target domain, domain generalization aims to learn representations that remain effective in unseen environments. Recent foundation-model approaches, such as CrossEarth, improve domain generalization through style diversification and multi-task representation learning (Gong et al., 2026). Data augmentation provides another strategy for increasing source-domain diversity. For example, Diffusion-Noise-based Augmentation combines diffusion-generated samples with foreground-background mixing to improve robustness under imbalanced remote sensing data (Wang et al., 2025). These methods mainly focus on improving the robustness of a single model, whereas our work studies how robust knowledge can be transferred from a large teacher to a compact student without propagating domain-specific shortcuts.

Knowledge Distillation for Semantic Segmentation. Knowledge distillation transfers knowledge from a high-capacity teacher to a compact student. Early methods match softened output distributions (Hinton et al., 2015), while segmentation-oriented approaches exploit spatial structures in dense predictions. Structured knowledge distillation captures pixel relations (Liu et al., 2019); attention transfer and channel-wise distillation align spatial activation patterns and channel distributions (Zagoruyko & Komodakis, 2017; Shu et al., 2021); and double similarity distillation models pixel- and category-level relations (Feng et al., 2021). Recent approaches further improve distillation under noisy supervision or varying learning difficulty. Boundary-aware Knowledge Distillation reduces unreliable annotation effects in remote-sensing segmentation (Sun et al., 2025), while Relative Difficulty Distillation adaptively guides pixel-wise transfer using teacher-student difficulty (Liang et al., 2024). Despite these advances, most methods optimize teacher-student agreement on source-domain samples, potentially transferring domain-specific teacher behaviors to students (Lukasik et al., 2022). Generalizable Knowledge Distillation improves out-of-domain segmentation by selectively transferring transferable knowledge from vision foundation models (Lv et al., 2026). In contrast, we consider two complementary sources of non-causal knowledge: domain-specific information transferred by the teacher and spurious cues acquired by the student. Our framework addresses both through teacher-side knowledge refinement and student-side intervention, while remaining compatible with existing distillation objectives.

Causal Learning. Causal representation learning aims to learn predictors based on stable, task-relevant factors rather than training-specific correlations. This perspective is important under visual domain shifts, where models may exploit background, texture, illumination, or acquisition artifacts as shortcuts (Zhou et al., 2021). Frequency-domain representations provide an effective space for manipulating appearance variations while preserving semantic structures. FACT improves domain generalization by perturbing Fourier statistics related to image style (Xu et al., 2021). Causal-Tune applies DCT-based feature manipulation to mine stable factors from vision foundation models for domain-generalized segmentation (Zhang et al., 2026), while CVGL introduces content-adaptive frequency intervention to reduce confounding signals in cross-view geo-localization (Ouyang & Zhu, 2025). Adversarial learning provides another mechanism for suppressing domain-dependent information. DANN learns domain-invariant representations through gradient reversal (Ganin et al., 2016), and MAD adversarially mines non-causal factors for unseen-domain detection (Xu et al., 2023). However, these methods mainly improve individual model robustness and do not consider how domain-specific knowledge is transferred or acquired during distillation. Our framework bridges causal robust learning and knowledge distillation by refining teacher knowledge before transfer and intervening in student learning.

3 Method

3.1 Causal Modeling

Figure 2(a) illustrates our causal view (Pearl et al., 2016). Let C denote semantic content, such as land-cover identity, object geometry, and spatial layout, and let D denote domain factors, includ-

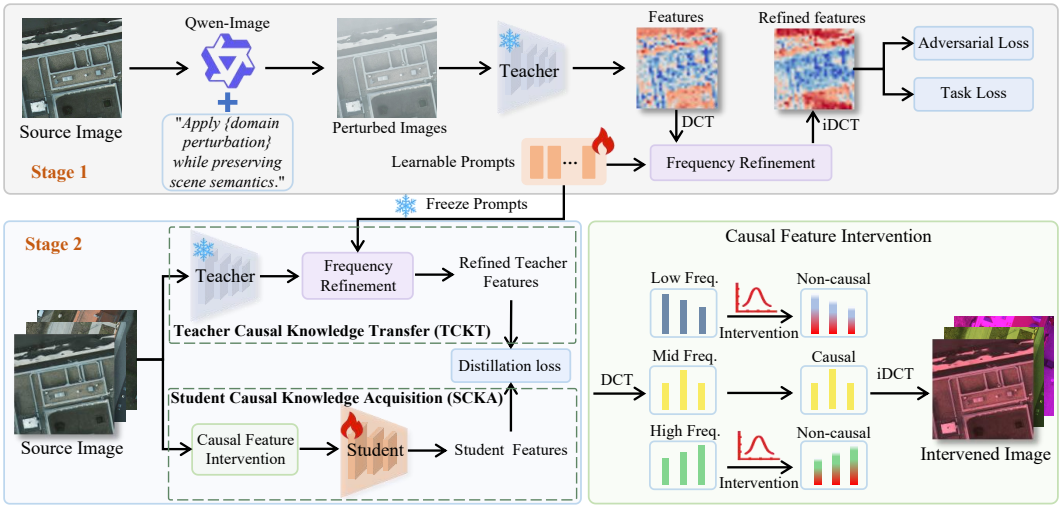


Figure 3: Overview of our CKD pipeline. Stage 1: TCKT learns frequency prompts from domain-perturbed views to suppress domain-sensitive teacher responses while preserving causal and task-relevant knowledge. Stage 2: The prompts are frozen to refine teacher features for distillation, while SCKA applies causal frequency intervention to guide the student to acquire causal knowledge.

ing sensor characteristics, weather, and noise. The source image X is affected by both C and D, while the segmentation label Y is determined only by C, following $C \rightarrow X \leftarrow D$ and $C \rightarrow Y$. Therefore, variations in D may change the appearance of X without altering its semantic content or pixel-wise labels. However, source-domain training may encourage models to exploit spurious correlations between D and predictions, leading to noncausal shortcut representations. A segmentation model extracts content-related features F_c and domain-related features F_d from X and predicts $\widehat{Y} = h(F_c, F_d)$. Here, F_c captures semantic information associated with C, whereas F_d captures noncausal domain information associated with D that may become shortcuts under source-domain training. Note that F_c and F_d denote functional roles rather than an explicit feature decomposition.

As shown in Figure 2(b), noncausal knowledge can enter the student through two routes. First, the teacher may encode domain-dependent responses F_d^T together with semantic knowledge F_c^T, causing conventional distillation to transfer noncausal information to the student. Second, source-domain supervision may drive the student to acquire domain-related features F_d^S as predictive shortcuts. Motivated by these two sources, Figure 2(c) introduces two complementary mechanisms. TCKT suppresses domain-dependent teacher responses F_d^T before distillation, while SCKA intervenes in student learning to reduce reliance on F_d^S. Together, they mitigate noncausal shortcuts and encourage stable representations under domain shifts.

3.2 Teacher Causal Knowledge Transfer (TCKT)

TCKT aims to reduce the transfer of domain-dependent noncausal knowledge from the teacher. Previous studies have shown that frequency-domain representations, particularly DCT-based features, provide an effective space for improving robustness under domain shifts (Xu et al., 2021; Zhang et al., 2026). Motivated by this observation, TCKT learns frequency prompts to identify and suppress domain-sensitive teacher responses. As illustrated in Figure 3, TCKT operates across two stages. Stage 1 learns frequency prompts from domain-perturbed views, while Stage 2 freezes the prompts to refine teacher features during distillation.

Specifically, in Stage 1, we first generate domain-perturbed views to expose domain-sensitive teacher responses. For each labeled source image (X, Y), we use Qwen-Image (Wu et al., 2025) in image-editing mode to generate a set of domain-perturbed views $\{X^{(d)}\}_{d\in\mathcal{D}}$ with different domain perturbations. The editing process changes appearance factors while preserving scene semantics. We then feed each domain-perturbed view $X^{(d)}$ into the frozen teacher to extract an intermediate feature $F_l^T \in \mathbb{R}^{B\times C_T\times H_l\times W_l}$ at layer l. The feature is transformed into the frequency domain using

a two-dimensional discrete cosine transform (DCT) (Ahmed et al., 1974):

$$S_l^T = \mathcal{D}_{\mathrm{DCT}}(F_l^T). \tag{1}$$

Instead of manually defining frequency bands, TCKT learns a bank of K frequency prompts $P = \{p_k\}_{k=1}^K$ shared across teacher layers to adaptively determine which frequency responses should be retained. For the channel vector $s_{l,u}^T$ at frequency bin u, we compute a continuous retention mask:

$$M_{l,u} = \sigma\left(\log \frac{1}{K} \sum_{k=1}^{K} \exp\left(\frac{\bar{p}_k^\top \bar{s}_{l,u}^T}{\tau_p}\right) + b\right), \tag{2}$$

where $\sigma(\cdot)$ denotes the sigmoid function, bars denote ℓ_2 normalization, τ_p is the prompt temperature, and b is a learnable retention bias. The refined teacher feature is then reconstructed as

$$\widetilde{F}_l^T = \mathcal{D}_{\mathrm{IDCT}}(M_l \odot S_l^T). \tag{3}$$

To learn effective frequency prompts, we jointly employ task and adversarial supervision during Stage 1. The segmentation loss encourages the refined teacher features to preserve task-relevant semantic information, while the adversarial objective suppresses domain-sensitive responses. Specifically, we introduce a lightweight domain discriminator Q, shared across the selected layers, with two heads trained using cross-entropy (CE) loss to predict the perturbation type d and severity r from each refined feature $\widetilde{F}_l^T$:

$$\mathcal{L}_{\mathrm{dom}} = \frac{1}{L} \sum_{l=1}^{L} \left[\mathrm{CE}\left(Q_{\mathrm{type}}(\widetilde{F}_l^T), d\right) + \lambda_{\mathrm{sev}}\, \mathrm{CE}\left(Q_{\mathrm{sev}}(\widetilde{F}_l^T), r\right)\right]. \tag{4}$$

Following domain-adversarial learning (Ganin et al., 2016), a gradient reversal layer (GRL) reverses the discriminator gradient passed to the frequency prompts. The discriminator learns to identify domain perturbations, whereas the prompts are optimized to remove frequency responses that reveal domain information. The complete TCKT objective is:

$$\mathcal{L}_{\mathrm{TCKT}} = \mathcal{L}_{\mathrm{seg}}^T(\widetilde{F}^T, Y) + \lambda_{\mathrm{adv}}\mathcal{L}_{\mathrm{dom}}^{\mathrm{GRL}} + \lambda_{\mathrm{bud}}\mathcal{L}_{\mathrm{bud}} + \lambda_{\mathrm{div}}\mathcal{L}_{\mathrm{div}}. \tag{5}$$

Here, $\mathcal{L}_{\mathrm{seg}}^T$ encourages the refined teacher features to preserve semantic prediction ability. Let $\overline{M}$ denote the average retention ratio over all frequency masks and selected layers. The budget term

$$\mathcal{L}_{\mathrm{bud}} = (\overline{M} - \rho)^2 \tag{6}$$

controls the overall retention ratio and prevents a trivial all-pass solution, where ρ is the target retention ratio. The diversity term

$$\mathcal{L}_{\mathrm{div}} = \frac{\|\bar{P}\bar{P}^\top - I\|_F^2}{K^2} \tag{7}$$

encourages diverse frequency prompts and prevents prompt collapse. Together, these objectives guide TCKT to preserve stable, task-relevant semantic information while suppressing domain-sensitive teacher responses. In Stage 2, the frozen prompts are applied to teacher features through Eqs. 1–3, producing refined teacher representations $\widetilde{F}^T$ as distillation targets.

3.3 Student Causal Knowledge Acquisition (SCKA)

In Stage 2, although TCKT suppresses domain-sensitive knowledge transferred from the teacher, the student may still acquire noncausal shortcuts from source-domain supervision. SCKA therefore introduces a content-adaptive frequency intervention (Ouyang & Zhu, 2025), which more strongly perturbs domain-sensitive low- and high-frequency components while relatively preserving structurally informative mid-frequency components, as illustrated in Figure 3.

For each source image X, we apply a two-dimensional DCT independently to its three channels to obtain the spectrum S_X. We then construct an image-dependent spectral weighting W_X based on image gradient magnitude and frequency location. Strong-gradient regions, which usually correspond to object boundaries and structural details, receive weaker perturbations, whereas smoother regions are perturbed more strongly. The intervened input is constructed as

$$X^{\mathrm{int}} = \mathrm{Norm}\left(\mathrm{clip}\left[\mathcal{D}_{\mathrm{IDCT}}\left(S_X \odot (1 + z\,\xi \odot W_X)\right)\right]\right), \tag{8}$$

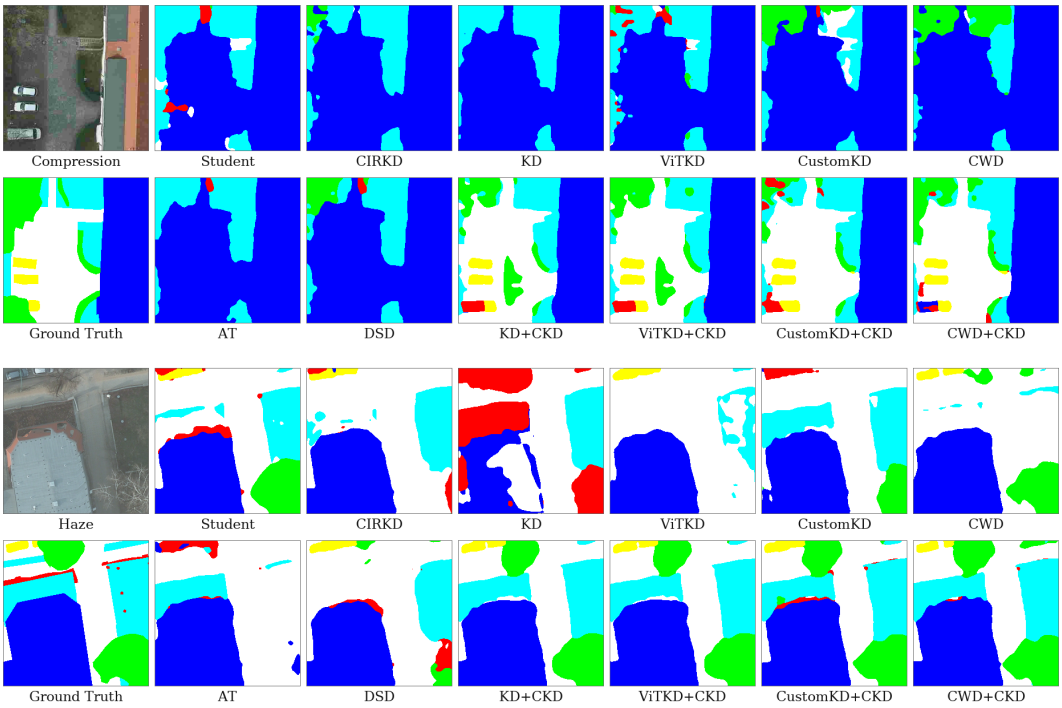


Figure 4: Qualitative comparison under domain shifts induced by Compression Artifacts (top two rows) and Haze (bottom two rows). We show the corrupted inputs, ground truth, and predictions from the student, conventional distillation methods, and their CKD-augmented counterparts.

where W_X adaptively controls the intervention strength while preserving task-relevant structural information, $z \sim \text{Bernoulli}(p)$ controls whether the intervention is applied, and $\xi_c \sim \mathcal{N}(0, \sigma^2)$ controls its strength for channel c. Since the intervention changes spectral appearance without altering spatial correspondence, the original segmentation label Y remains valid.

During distillation, the frozen teacher processes the clean image X, while the student receives the intervened image X^{int}. The frozen TCKT prompts refine the teacher features following Eq. 3, yielding $\widetilde{F}^T(X)$ as the distillation target. The student is optimized by

$$\mathcal{L}_{\text{SCKA}} = \mathcal{L}^S_{\text{seg}}(X^{\text{int}}, Y) + \lambda_{\text{KD}} \mathcal{L}_{\text{KD}} \left(F^S(X^{\text{int}}), \widetilde{F}^T(X) \right), \tag{9}$$

where $\mathcal{L}_{\text{KD}}$ can be instantiated with different distillation objectives. By learning from frequency-intervened inputs while matching the refined teacher representation, the student is encouraged to capture stable task-relevant information rather than domain-dependent frequency shortcuts.

4 Experiments

4.1 Experimental Setup

Datasets. We train all experiments on the ISPRS Potsdam dataset (Rottensteiner et al., 2012), which serves as the source domain. The source-domain split contains 3,456 training, 400 validation, and 2,016 test images with a spatial resolution of 512×512. We evaluate the models on the source domain and eight corruption-induced domain shifts from the REOBench benchmark (Li et al., 2025). The domain-shift evaluation includes eight corruption types: Brightness Contrast, Cloud, Compression Artifacts, Gaussian Blur, Gaussian Noise, Haze, Motion Blur, and Salt and Pepper Noise, covering diverse environmental, sensor, and acquisition variations. Each domain shift contains five severity levels, with 2,016 images per severity level. These domain-shift data are used only for evaluation and are never involved in source-domain training. They are also distinct from the domain perturbations generated in TCKT, which are used only for learning frequency prompts.

Table 1: Foreground-class mIoU (%) on the clean Potsdam source domain and eight corruption-induced domain shifts. Avg. denotes the mean over the eight domain-shift conditions. Results marked with † are taken from REOBench (Li et al., 2025). The "+CKD" denotes the corresponding distillation method augmented with our CKD.

Method	Backbone	Source Domain	Brightness Contrast	Cloud	Compression Artifacts	Gauss Blur	Gauss Noise	Haze	Motion Blur	Salt Pepper	Avg.
MIM- and CL-based methods											
SatMAE† (NeurIPS 2022)	ViT-L	59.51	50.18	37.39	48.23	57.89	41.83	44.95	57.52	36.07	46.76
ScaleMAE† (ICCV 2023)	ViT-L	68.92	64.37	65.43	49.96	64.86	54.89	63.89	64.49	52.12	60.00
RVSA† (TGRS 2023)	ViT-B	69.82	64.71	65.67	47.99	66.89	48.89	61.52	65.25	45.26	58.27
SatMAE++† (CVPR 2024)	ViT-L	62.68	53.91	59.06	49.74	60.38	44.32	48.80	60.44	39.45	52.01
RemoteCLIP† (TGRS 2024)	ViT-B	50.28	42.32	45.27	39.33	50.26	48.46	36.61	50.06	46.91	44.90
RemoteCLIP† (TGRS 2024)	ViT-L	56.69	54.51	52.73	43.24	50.82	45.12	50.47	51.82	38.98	48.46
GeoRSCLIP† (TGRS 2024)	ViT-B	51.44	42.89	46.41	40.37	51.35	49.79	38.56	51.15	48.24	46.10
GeoRSCLIP† (TGRS 2024)	ViT-L	56.81	54.97	52.53	43.28	50.49	42.70	49.41	51.36	36.98	47.72
Reference networks											
Teacher	ViT-L	80.93	77.75	77.90	72.88	78.71	73.09	76.02	76.23	72.48	75.63
Student	ViT-B	75.73	65.98	69.91	67.40	69.95	69.00	63.79	69.87	68.75	68.08
Conventional distillation											
KD (NeurIPS 2015)	ViT-B	76.06	63.61	67.21	65.34	67.21	67.11	60.56	67.35	66.76	65.64
AT (ICLR 2017)	ViT-B	76.38	64.41	67.61	65.65	68.38	67.18	61.66	68.39	66.92	66.28
DSD (TIP 2021)	ViT-B	76.41	64.00	67.83	66.23	67.72	66.25	61.26	67.96	65.10	65.79
CWD (ICCV 2021)	ViT-B	78.17	66.32	70.17	68.06	71.04	67.10	64.72	70.48	66.40	68.04
CIRKD (CVPR 2022)	ViT-B	77.68	63.10	67.18	65.47	67.46	64.79	60.78	66.63	64.40	64.98
ViTKD (CVPRW 2024)	ViT-B	76.12	60.76	64.32	62.11	65.44	64.74	58.49	64.64	64.46	63.12
CustomKD (CVPR 2025)	ViT-B	76.71	63.96	67.81	65.50	67.12	67.75	61.00	67.99	67.03	66.02
Generalizable distillation											
GKD (CVPR 2026)	ViT-B	78.99	**78.19**	76.97	68.84	**78.20**	72.12	77.62	77.60	68.24	74.72
Causal distillation (ours)											
KD+CKD (NeurIPS 2015)	ViT-B	77.52	69.30	72.97	71.58	73.56	72.66	67.45	72.92	72.41	71.61
CWD+CKD (ICCV 2021)	ViT-B	78.35	69.93	74.73	72.82	73.78	73.96	67.75	74.26	72.08	72.41
ViTKD+CKD (CVPRW 2024)	ViT-B	77.36	69.16	72.64	71.31	73.46	73.09	66.59	72.85	72.37	71.43
CustomKD+CKD (CVPR 2025)	ViT-B	77.04	69.12	72.93	71.65	73.38	72.42	67.37	73.22	70.74	71.35
GKD+CKD (CVPR 2026)	ViT-B	**79.33**	77.95	**78.30**	**74.80**	78.05	**76.50**	**79.30**	**79.20**	**72.94**	**77.13**

Models and compared methods. We use DINOv3 ViT-L as the teacher and DINOv3 ViT-B as the student (Siméoni et al., 2025), both equipped with a Mask2Former decoder (Cheng et al., 2022). We consider representative KD methods, including KD (Hinton et al., 2015), AT (Zagoruyko & Komodakis, 2017), CWD (Shu et al., 2021), DSD (Feng et al., 2021), CIRKD (Yang et al., 2022), CustomKD (Lee et al., 2025), ViTKD (Yang et al., 2024), and GKD (Lv et al., 2026). For comparison, we also include masked image modeling (MIM)-based remote-sensing methods, including SatMAE (Cong et al., 2022), ScaleMAE (Reed et al., 2023), RVSA (Wang et al., 2023), and SatMAE++ (Noman et al., 2024), as well as contrastive learning (CL)-based methods, RemoteCLIP (Liu et al., 2024) and GeoRSCLIP (Zhang et al., 2024). To evaluate the generality of CKD, we integrate it with different distillation objectives and compare the resulting variants. In the following experiments, the "+CKD" suffix denotes a distillation method augmented with CKD, allowing direct comparison with its original counterpart.

Implementation details. CKD is trained in two stages. First, the teacher is frozen and the TCKT prompt bank ($K = 16$) is optimized for 200 epochs on domain-perturbed images using segmentation and adversarial supervision. The prompts and discriminator are trained with AdamW using learning rates of 1×10^{-3} and 1×10^{-4}, respectively, with $\lambda_{\text{adv}} = 0.3$, $\lambda_{\text{bud}} = 0.05$, $\lambda_{\text{div}} = 0.01$, and $\rho = 0.1$. The learned prompts are then frozen, and the student is further trained for 200 epochs with SCKA using the segmentation loss and the selected distillation objective. We use AdamW with a learning rate of 1×10^{-4}, weight decay of 0.05, and a polynomial learning-rate schedule with power 0.9. For SCKA, we set $p = 0.8$ and $\sigma = 0.5$.

4.2 Qualitative Results

Figure 4 presents qualitative comparisons under two representative corruption-induced domain shifts: Compression Artifacts and Haze. Conventional distillation methods exhibit degraded segmentation quality under these challenging conditions, including fragmented predictions, inaccurate object boundaries, and missing small structures. Under Haze, reduced contrast further increases con-

Table 2: Component ablation of TCKT and SCKA. Results are reported as mean mIoU over eight domain-shift conditions.

TCKT	SCKA	KD	CWD	ViTKD	CustomKD
×	×	65.64	68.04	63.12	66.02
✓	×	69.29	71.13	70.07	69.59
×	✓	68.57	70.54	69.34	68.78
✓	✓	**71.61**	**72.41**	**71.43**	**71.35**

Table 3: Effect of the number of domain perturbation types. Results are reported as mean mIoU over eight domain-shift conditions.

Types	Teacher	KD+CKD	CWD+CKD	ViTKD+CKD	CustomKD+CKD
1	78.53	70.78	70.60	**71.49**	69.98
3	78.65	71.61	**72.41**	71.43	**71.35**
5	**78.95**	**71.88**	71.98	71.25	70.24
7	78.91	69.82	71.56	71.41	70.26

fusion between buildings, vegetation, and background regions. In contrast, CKD-enhanced variants generate more complete semantic regions with clearer boundaries and better preservation of small objects across both domain shifts. These improvements indicate that CKD reduces the influence of domain-sensitive appearance cues by refining teacher knowledge through TCKT and encouraging the student to learn more stable semantic representations through SCKA.

4.3 Quantitative Results

Table 1 reveals a substantial robustness gap between the source domain and corruption-induced domain shifts. Although remote-sensing models achieve strong performance on the ISPRS Potsdam source domain, their performance degrades substantially under REOBench domain shifts. Conventional distillation methods generally improve source-domain accuracy but provide limited robustness gains under domain shifts, suggesting that standard knowledge transfer may preserve or amplify domain-dependent noncausal cues. Even GKD, which improves robustness compared with conventional objectives, still exhibits noticeable performance degradation under domain shifts. In contrast, CKD consistently improves different distillation objectives by addressing both sources of noncausal knowledge: TCKT reduces the transfer of domain-sensitive teacher responses, while SCKA mitigates the acquisition of domain-dependent shortcuts during student learning. Across different distillation objectives, CKD improves robustness under diverse domain shifts while maintaining or improving source-domain performance, demonstrating its effectiveness for robust knowledge distillation in remote sensing semantic segmentation.

Table 4: Foreground-class mIoU (%) of DeepLabV3 models on the ISPRS Potsdam source domain and eight corruption-induced domain shifts. The "+CKD" denotes the corresponding distillation method augmented with our CKD.

Method	Backbone	Source Domain	Brightness Contrast	Cloud	Compression Artifacts	Gauss Blur	Gauss Noise	Haze	Motion Blur	Salt Pepper	Avg.
Reference networks											
Teacher	ResNet-101	80.61	49.43	76.66	68.45	76.75	68.99	31.56	76.03	62.98	63.86
Student	ResNet-18	76.79	40.67	72.13	60.96	72.88	61.08	23.64	72.29	56.82	57.56
Conventional distillation											
AT (ICLR 2017)	ResNet-18	76.52	40.72	71.29	61.28	72.25	63.66	29.04	72.75	57.06	57.25
DSD (TIP 2021)	ResNet-18	76.86	42.87	72.85	62.64	72.72	62.81	33.79	72.30	55.18	58.14
CWD (ICCV 2021)	ResNet-18	78.68	44.79	74.02	64.28	74.40	66.21	34.15	71.71	58.74	59.79
CIRKD (CVPR 2022)	ResNet-18	76.94	41.73	72.71	61.59	73.06	62.68	33.02	71.92	57.84	58.07
Causal distillation (ours)											
AT+CKD (ICLR 2017)	ResNet-18	77.31	42.24	75.66	67.69	75.37	69.52	31.10	76.53	63.88	62.75
DSD+CKD (TIP 2021)	ResNet-18	77.44	45.02	77.24	69.03	76.81	69.37	36.11	76.46	62.75	64.10
CWD+CKD (ICCV 2021)	ResNet-18	**79.02**	**47.05**	**79.36**	**71.48**	**78.52**	**73.11**	**37.02**	**76.84**	**67.39**	**66.35**
CIRKD+CKD (CVPR 2022)	ResNet-18	77.67	43.51	76.88	68.26	77.21	69.27	35.17	76.11	65.08	63.93

4.4 Ablation Studies

Effect of components. Table 2 evaluates TCKT and SCKA across four distillation objectives. Both improve corruption robustness, with their combination performing best throughout. For KD, TCKT and SCKA raise the corruption average from 65.64 to 69.29 and 68.57, respectively, and together reach 71.61. The full model also improves ViTKD from 63.12 to 71.43, CustomKD from 66.02 to 71.35, and CWD from 68.04 to 72.41. These gains confirm the complementary effects of teacher-side refinement and student-side intervention across distillation objectives.

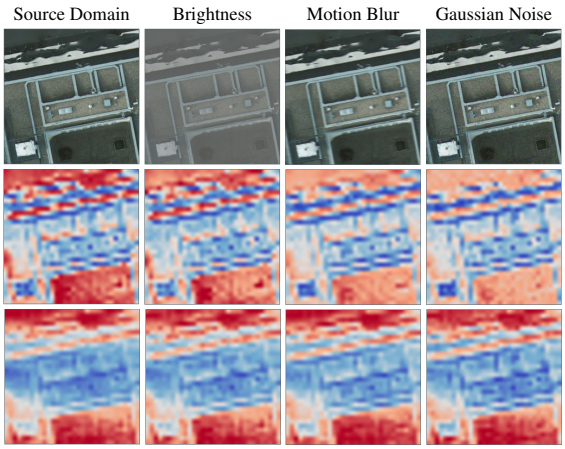


Figure 5: Visualization of TCKT-refined teacher features on the ISPRS Potsdam source domain and three corruption-induced domain shifts. Rows show the input image, original teacher feature, and TCKT-refined feature, respectively.

Table 5: Effect of the number of learnable frequency prompts on the ISPRS Potsdam source domain and eight corruption-induced domain shifts in terms of foreground-class mIoU (%). The "+CKD" denotes the corresponding method augmented with our CKD.

Method	Prompt Number			
	$K = 8$	$K = 12$	$K = 16$	$K = 20$
Source Domain				
Teacher	81.25	81.38	81.41	**81.49**
KD+CKD	76.65	77.11	77.52	**77.55**
CWD+CKD	77.44	77.83	**78.35**	78.31
ViTKD+CKD	76.47	76.98	**77.36**	77.33
CustomKD+CKD	76.31	76.62	77.04	**77.06**
Corruption Avg.				
Teacher	76.22	77.61	**78.65**	78.44
KD+CKD	66.84	69.18	71.61	**71.65**
CWD+CKD	67.22	70.03	**72.41**	72.36
ViTKD+CKD	66.39	69.08	**71.43**	71.41
CustomKD+CKD	66.56	68.71	71.35	**71.38**

Effect of perturbation diversity. Table 3 studies how the number of domain perturbation types used for generating training images affects TCKT prompt learning. The generated perturbation types are different from the eight corruption-induced domain shifts used for evaluation. Under the default setting with three perturbation types, the refined teacher model improves from 75.63 to 78.65 compared with the original teacher model reported in Table 1. This setting provides sufficient domain diversity to extract stable content knowledge without introducing unnecessary training complexity.

Effect of different backbones. We further evaluate CKD with a DeepLabV3-ResNet-101 teacher and a DeepLabV3-ResNet-18 student to examine its robustness across different architectures. As shown in Table 4, CKD improves different distillation methods under diverse corruption-induced domain shifts, with clear gains on compression, noise, and other challenging conditions. These results demonstrate that CKD provides robust performance beyond a specific backbone family.

Effect of prompt number. Table 5 investigates the effect of the number of learnable frequency prompts. Increasing the prompt number initially improves robustness under domain shifts across different distillation objectives, while further increasing it brings marginal gains or slight degradation. We therefore adopt $K = 16$ as the default setting. Figure 5 visualizes the effect of the learned prompts. Compared with the original teacher features, TCKT-refined features exhibit consistent spatial responses, with reduced domain-shift-induced variations. These results indicate that TCKT effectively suppresses domain-sensitive responses while preserving stable task-relevant structures.

5 Conclusion

In this paper, we investigate the limited robustness of knowledge distillation under domain shifts in remote sensing semantic segmentation and identify two sources of domain-dependent noncausal knowledge: noncausal responses transferred from the teacher and noncausal shortcuts acquired by the student from source-domain supervision. To address these issues, we propose Causal Knowledge Distillation (CKD), which combines Teacher Causal Knowledge Transfer (TCKT) and Student Causal Knowledge Acquisition (SCKA) to regulate teacher knowledge transfer and student learning, respectively. CKD can be integrated with different distillation objectives without modifying the student architecture or introducing additional inference cost. Experiments on the ISPRS Potsdam source-domain dataset and multiple corruption-induced domain shifts demonstrate that CKD improves the robustness of different distillation objectives under domain shifts while maintaining strong source-domain performance. These results highlight the potential of CKD for improving the reliability of lightweight remote sensing models in practical deployment scenarios, including resource-constrained onboard applications.

AI USE STATEMENT

Generative AI tools were used to assist with code implementation and debugging, literature search, and manuscript editing. They were not used to generate research ideas, experimental data, results, or scientific claims. All AI-assisted outputs were reviewed, verified, and edited by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses the publicly available ISPRS Potsdam dataset and the REOBench evaluation protocol. The ISPRS Potsdam dataset is used under its original terms, and REOBench is used as released by its authors. This work does not involve human subjects, personally identifiable information, or sensitive personal data. We confirm that this work complies with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed experimental settings, hyperparameters, and evaluation protocols in the paper. The code, configuration files, domain perturbation generation pipeline, evaluation scripts, and instructions for reproducing the reported results will be made publicly available upon acceptance.

REFERENCES

Nasir Ahmed, T. Natarajan, and Kamisetty R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974.

Gaétan Bahl, Lionel Daniel, Matthieu Moretti, and Florent Lafarge. Low-power neural networks for semantic segmentation of satellite images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2469–2476. IEEE, 2019.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1280–1289. IEEE, 2022.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. DeepGlobe 2018: A challenge to parse the Earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 172–181. IEEE, 2018.

Yingchao Feng, Xian Sun, Wenhui Diao, Jihao Li, and Xin Gao. Double similarity distillation for semantic image segmentation. *IEEE Transactions on Image Processing*, 30:5363–5376, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Ziyang Gong, Zhixiang Wei, Di Wang, Xiaoxing Hu, Xianzheng Ma, Hongruixuan Chen, Yuru Jia, Yupeng Deng, Zhenming Ji, Xiangwei Zhu, Xue Yang, Naoto Yokoya, Jing Zhang, Bo Du, Junchi Yan, and Liangpei Zhang. CrossEarth: Geospatial vision foundation model for domain generalizable remote sensing semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(5):5147–5164, 2026. doi: 10.1109/TPAMI.2025.3649001.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Wei Huang, Zhiliang Peng, Li Dong, Furu Wei, Jianbin Jiao, and Qixiang Ye. Generic-to-specific distillation of masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15996–16005, 2023.

Jungsoo Lee, Debasmit Das, Munawar Hayat, Sungha Choi, Kyuwoong Hwang, and Fatih Porikli. CustomKD: Customizing large vision foundation for edge model improvement via knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25176–25186. IEEE, 2025.

Xiang Li, Yong Tao, Siyuan Zhang, Siwei Liu, Zhitong Xiong, Chunbo Luo, Lu Liu, Mykola Pechenizkiy, Xiaoxiang Zhu, and Tianjin Huang. REOBench: Benchmarking robustness of Earth observation foundation models. *Advances in Neural Information Processing Systems*, 38, 2025.

Dong Liang, Yue Sun, Yun Du, Songcan Chen, and Sheng-Jun Huang. Relative difficulty distillation for semantic segmentation. *Science China Information Sciences*, 67(9):192105, 2024.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. RemoteCLIP: A vision-language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2604–2613, 2019.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Teacher's pet: Understanding and mitigating biases in distillation. *Transactions on Machine Learning Research*, 2022.

Chonghua Lv, Dong Zhao, Shuang Wang, Dou Quan, Ning Huyan, Nicu Sebe, and Zhun Zhong. Generalizable knowledge distillation from vision foundation models for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27811–27819, 2024.

Songsong Ouyang and Yingying Zhu. CVGL: Causal learning and geometric topology. *Advances in Neural Information Processing Systems*, 38:78247–78266, 2025.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4088–4099. IEEE, 2023.

Caleb Robinson, Le Hou, Kolya Malkin, Rachel Soobitsky, Jacob Czawlytko, Bistra Dilkina, and Nebojsa Jojic. Large-scale high-resolution land-cover mapping with multi-resolution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12718–12727. IEEE, 2019.

Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1:293–298, 2012.

Jacob Shermeyer, Daniel Hogan, Jason Brown, Adam Van Etten, Nicholas Weir, Fabio Pacifici, Ronny Hänsch, Alexei Bastidas, Scott Soenen, Todd Bacastow, et al. SpaceNet 6: Multi-sensor all-weather mapping dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 768–777. IEEE, 2020.

Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5291–5300. IEEE, 2021.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Yue Sun, Dong Liang, Shao-Yuan Li, Songcan Chen, and Sheng-Jun Huang. Handling noisy annotation for remote sensing semantic segmentation via boundary-aware knowledge distillation. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–20, 2025.

Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran Associates, Inc., 2021.

Qianqian Wang, Haibo Ye, Dong Liang, and Sheng-Jun Huang. Diffusion-noise-based augmentation for long-tailed remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Mingjun Xu, Lingyun Qin, Weijie Chen, Shiliang Pu, and Lei Zhang. Multi-view adversarial discriminator: Mine the non-causal factors for object detection in unseen domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8103–8112. IEEE, 2023.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A Fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14378–14387. IEEE, 2021.

Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12309–12318. IEEE, 2022.

Zhendong Yang, Zhe Li, Ailing Zeng, Zexian Li, Chun Yuan, and Yu Li. ViTKD: Feature-based knowledge distillation for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1379–1388. IEEE, 2024.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

Yin Zhang, Yongqiang Zhang, Yaoyue Zheng, Bogdan Raducanu, and Dan Liu. Causal-Tune: Mining causal factors from vision foundation models for domain generalized semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12916–12924, 2026.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. RS5M and GeoRSCLIP: A large-scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–23, 2024.

Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. Examining and combating spurious features under distribution shift. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 12857–12867. PMLR, 2021.

SUPPLEMENTARY MATERIAL

This supplementary material provides additional methodological, implementation, and visual evidence that complements the main paper. It is organized as follows:

§A: Domain Shifts in Remote Sensing
§B: Additional Method and Training Details
§C: Domain-Perturbed Image Generation
§D: Cross-Dataset Generalization on Vaihingen
§E: Additional Visual Results
§F: Feature-Space Evidence of Stable Student Representations

A DOMAIN SHIFTS IN REMOTE SENSING

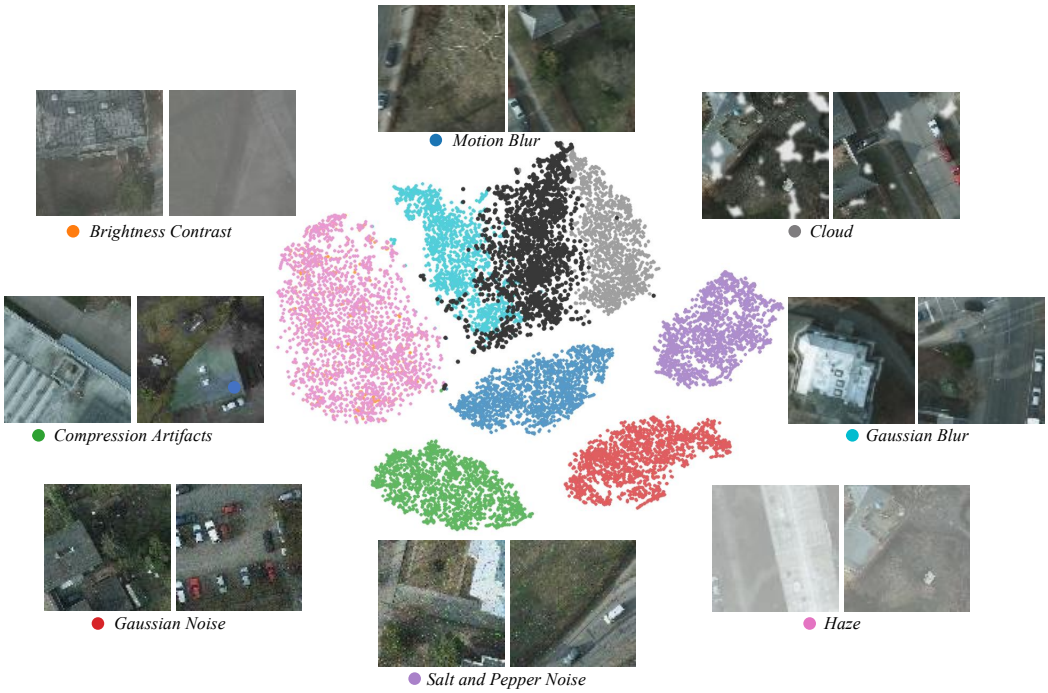


Figure S1: t-SNE visualization of feature distributions extracted from the final layer of the DINOv3 ViT-L teacher on the Potsdam source domain (Rottensteiner et al., 2012) and eight corruption-induced shifted domains from REOBench (Li et al., 2025). • denotes samples from the source domain, while colored points represent samples from different shifted domains. The evident distribution discrepancy between the source domain and corrupted domains demonstrates substantial domain shifts in the teacher feature space, indicating that foundation model representations are affected by domain-specific variations under distribution shifts.

Domain shift is a critical challenge for reliable remote sensing interpretation, particularly for models deployed on satellites and other resource-constrained platforms. During real-world acquisition, remote sensing images collected by different sensors, at different times, or across different geographic regions may exhibit substantial variations due to sensor spectral responses, spatial resolution, illumination conditions, atmospheric and weather effects, platform motion, data compression, and sensor noise (Gong et al., 2026; Li et al., 2025). These variations can significantly alter the observed image appearance while preserving the underlying land-cover semantics. Consequently, models trained on a single source domain may inadvertently rely on domain-dependent visual cues that are effective during training but unstable under unseen deployment conditions.

This issue is particularly challenging for satellite-based applications. Unlike conventional terrestrial systems, onboard models must operate under strict constraints on computation, memory, and power consumption, making frequent target-domain annotation collection and model adaptation difficult or even infeasible. Therefore, improving robustness under domain shifts is essential for achieving consistent and reliable land-cover mapping across different sensors, geographic regions, and acquisition environments, which is critical for practical remote sensing deployment.

Following the REOBench evaluation protocol (Li et al., 2025), we evaluate model robustness under eight corruption-induced domain shifts defined by REOBench based on the ISPRS Potsdam source-domain test set. These domain shifts include Brightness Contrast, Cloud, Compression Artifacts, Gaussian Blur, Gaussian Noise, Haze, Motion Blur, and Salt and Pepper Noise. Each shift simulates representative variations in remote sensing acquisition while preserving scene structures and semantic labels, providing a controlled setting to evaluate whether models capture stable semantic representations or rely on appearance-dependent cues.

Figure S1 visualizes representative samples from these domain shifts and their corresponding feature distributions. The surrounding panels show two representative examples from each domain-shift condition, illustrating the diverse appearance variations introduced by different acquisition factors. The feature distributions further reveal substantial shifts between the source domain and corruption-induced domain shifts in the DINOv3 ViT-L teacher representation space. Although the scene content and semantic labels remain unchanged, different shift types produce distinct feature displacements or separated clusters, indicating that teacher representations remain sensitive to domain-dependent appearance variations.

These observations highlight an important limitation of conventional knowledge distillation under domain shifts. A powerful pretrained teacher may still encode domain-dependent information caused by acquisition variations, which can be unintentionally transferred to the student through standard distillation. Meanwhile, source-domain supervision may further encourage the student to capture unstable appearance cues as shortcuts. Therefore, maintaining strong source-domain performance alone is insufficient for reliable deployment; a robust distillation framework should regulate both the knowledge transferred from the teacher and the learning process of the student. This motivates our CKD framework, where TCKT suppresses domain-dependent teacher responses during knowledge transfer and SCKA encourages the student to acquire stable representations.

B Additional Method and Training Details

CKD is optimized in two stages. In Stage 1, the DINOv3 ViT-L teacher remains frozen, and TCKT learns frequency prompts from intermediate teacher features extracted from blocks 7, 11, 15, and 23. Each feature is transformed into the frequency domain along spatial dimensions using DCT, adaptively modulated by a shared bank of 16 learnable frequency prompts, and reconstructed using inverse DCT. Segmentation supervision encourages the retained responses to preserve task-relevant semantic information, while a shared domain discriminator with a gradient reversal layer guides the prompts to suppress responses predictive of perturbation type and severity. Only the frequency prompts and discriminator are optimized in this stage. After Stage 1, the discriminator is removed and the learned prompts are frozen.

In Stage 2, the frozen teacher processes the source-domain image, while the student receives its frequency-intervened counterpart. The intervention is applied with probability $p = 0.8$, using independent channel-wise Gaussian perturbations controlled by $\sigma = 0.5$. Since the intervention preserves spatial correspondence, the original segmentation labels remain unchanged. The frozen TCKT prompts are applied to refine teacher features, which are then used as distillation targets for the selected objective. We evaluate CKD with KD, ViTKD, CustomKD, and CWD, where only the corresponding distillation loss is replaced. CKD is not limited to these objectives and can be extended to other compatible distillation frameworks.

Both TCKT and SCKA are used only during training. During inference, the teacher, frequency prompt bank, domain discriminator, and frequency intervention are removed, and only the original student network is retained. Therefore, CKD introduces no additional inference parameters or computational overhead, making it suitable for deployment on resource-constrained remote sensing platforms. Algorithm 1 summarizes the complete two-stage training procedure.

Algorithm 1 Two-stage training procedure of CKD

Require: Source-domain dataset $\mathcal{D}_s = \{(X, Y)\}$; frozen teacher T; student S; frequency prompt bank P; domain discriminator Q; distillation objective $\mathcal{L}_{\mathrm{KD}}$
Ensure: Trained student S

Stage 1: Frequency prompt learning
Freeze teacher T
for each training sample (X, Y) **do**
  Generate a domain-perturbed view $X^{(d)}$
  Extract teacher features $\{F_l^T\}$ from selected blocks
  for each selected block l **do**
    Transform features: $S_l^T \leftarrow \mathcal{D}_{\mathrm{DCT}}(F_l^T)$
    Generate retention mask M_l with prompt bank P
    Refine features: $\widetilde{F}_l^T \leftarrow \mathcal{D}_{\mathrm{IDCT}}(M_l \odot S_l^T)$
  end for
  Compute semantic preservation loss and adversarial loss $\mathcal{L}_{\mathrm{TCKT}}$
  Update prompt bank P and discriminator Q
end for
Freeze prompt bank P and discard discriminator Q

Stage 2: Causal knowledge distillation
Freeze teacher T and prompt bank P
for each training sample (X, Y) **do**
  Generate frequency-intervened input X^{int}
  Extract teacher features $F^T(X)$ from the source-domain image
  Refine teacher features with frozen prompts: $\widetilde{F}^T(X)$
  Extract student features $F^S(X^{\mathrm{int}})$
  Compute segmentation and distillation losses:

$$\mathcal{L}_{\mathrm{SCKA}} = \mathcal{L}_{\mathrm{seg}}^S(X^{\mathrm{int}}, Y) + \lambda_{\mathrm{KD}}\mathcal{L}_{\mathrm{KD}}(F^S(X^{\mathrm{int}}), \widetilde{F}^T(X))$$

  Update student S
end for
return trained student S

C Domain-Perturbed Image Generation

The domain-perturbed views used in TCKT are generated offline using Qwen-Image in image-editing mode (Wu et al., 2025). We use snow, fog, and rain as the default perturbation types. For each source-domain image, the editing instruction modifies only weather conditions and imaging appearance while explicitly preserving the top-down viewpoint, scene layout, object locations, boundaries, and image resolution. The generated views therefore remain spatially aligned with the original images and can directly reuse the original pixel-level annotations.

These generated views are used only for learning TCKT frequency prompts. They are not included in the REOBench evaluation set and are not used to report domain-shift robustness results. Conversely, the eight REOBench corruption types are exclusively reserved for evaluation and are never used during prompt learning. This separation prevents overlap between the perturbations used to expose domain-sensitive teacher responses and the domain shifts used to assess robustness.

Table 3 further investigates the effect of perturbation diversity on frequency prompt learning. Using too few perturbation types provides limited variation for identifying domain-sensitive responses. However, increasing the number of perturbation types does not continuously improve performance, as excessive diversity may introduce unnecessary or less representative appearance variations. Based on this trade-off, we use three perturbation types by default, which provide sufficient domain diversity while maintaining stable prompt learning.

Table S1: Cross-dataset robustness evaluation on the Vaihingen target domain. Foreground-class mIoU (%) is reported for models trained on the ISPRS Potsdam source domain and directly evaluated on the Vaihingen test set without target-domain training or adaptation. The "+CKD" suffix denotes the corresponding distillation method augmented with CKD.

Method	Backbone	mIoU
Reference networks		
Teacher	ViT-L	72.26
Student	ViT-B	55.08
Conventional distillation		
KD (NeurIPS 2015)	ViT-B	55.16
AT (ICLR 2017)	ViT-B	54.94
DSD (TIP 2021)	ViT-B	52.93
CWD (ICCV 2021)	ViT-B	52.27
CIRKD (CVPR 2022)	ViT-B	52.48
ViTKD (CVPRW 2024)	ViT-B	53.62
CustomKD (CVPR 2025)	ViT-B	54.39
Causal distillation (ours)		
KD+CKD (NeurIPS 2015)	ViT-B	57.73
CWD+CKD (ICCV 2021)	ViT-B	**60.16**
ViTKD+CKD (CVPRW 2024)	ViT-B	<u>58.06</u>
CustomKD+CKD (CVPR 2025)	ViT-B	56.60

D Cross-Dataset Robustness Evaluation on Vaihingen

The REOBench experiments on Potsdam evaluate robustness against corruption-induced domain shifts within the same geographic dataset. To further assess cross-dataset domain robustness, we train all models on the ISPRS Potsdam source domain and directly evaluate them on the unseen Vaihingen dataset (Rottensteiner et al., 2012) without target-domain training or adaptation. We use the shared six-class label space and the same foreground-class mIoU metric as in Potsdam, and report performance on the native Vaihingen test set for each model. As shown in Table S1, CKD consistently improves all four paired distillation objectives on the unseen Vaihingen domain. The mIoU increases from 55.16 to 57.73 for KD, from 53.62 to 58.06 for ViTKD, from 54.39 to 56.60 for CustomKD, and from 52.27 to 60.16 for CWD. Among these methods, CWD+CKD achieves the highest compact-model performance with 60.16 mIoU, improving its conventional counterpart by 7.89 percentage points. These results complement the Potsdam corruption-induced domain-shift experiments by demonstrating that CKD also improves robustness when transferring across geographically distinct remote sensing datasets without target-domain adaptation.

E Additional Visual Results

Figures S2–S9 extend the qualitative comparison in Figure 4 to all eight REOBench corruption-induced domain shifts. Each domain shift changes the observed image appearance while preserving the underlying scene content and ground-truth semantics. Therefore, the resulting prediction errors mainly reflect model sensitivity to domain-dependent appearance variations rather than semantic changes. Across these domain shifts, conventional distillation methods frequently suffer from fragmented semantic regions, missing small objects, inaccurate boundaries, and spurious predictions. Such degradation is particularly evident when image brightness, contrast, texture, visibility, or local details differ from the source domain. In contrast, CKD-enhanced variants consistently produce more complete semantic regions, sharper object boundaries, and more stable predictions across diverse domain shifts. These qualitative results indicate that CKD reduces the reliance on domain-sensitive appearance cues by refining teacher knowledge through TCKT and guiding student learning through SCKA. The consistent improvements across different corruption-induced domain shifts further agree with the quantitative robustness gains reported in Table 1.

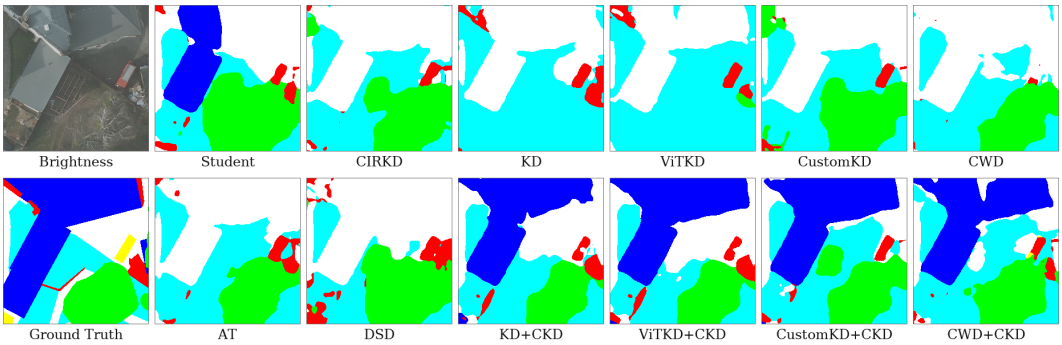


Figure S2: Qualitative comparison under the domain shift induced by Brightness Contrast.

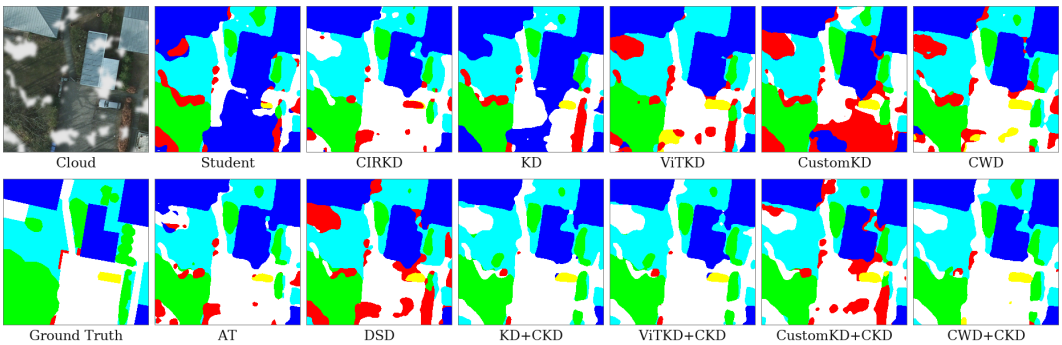


Figure S3: Qualitative comparison under the domain shift induced by Cloud.

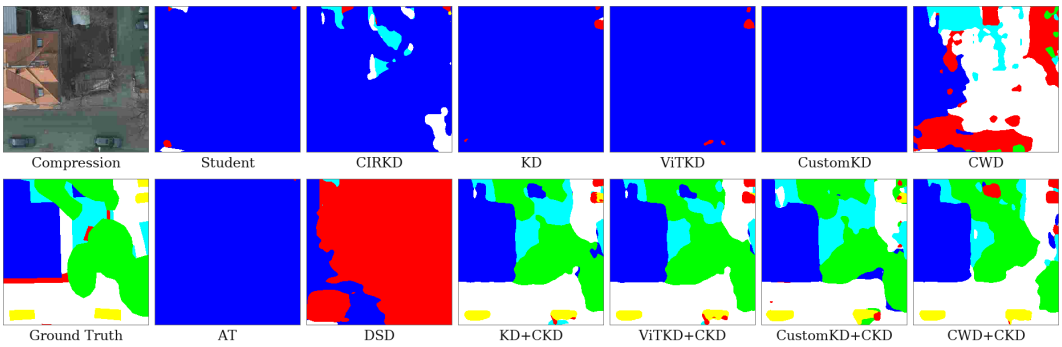


Figure S4: Qualitative comparison under the domain shift induced by Compression Artifacts.

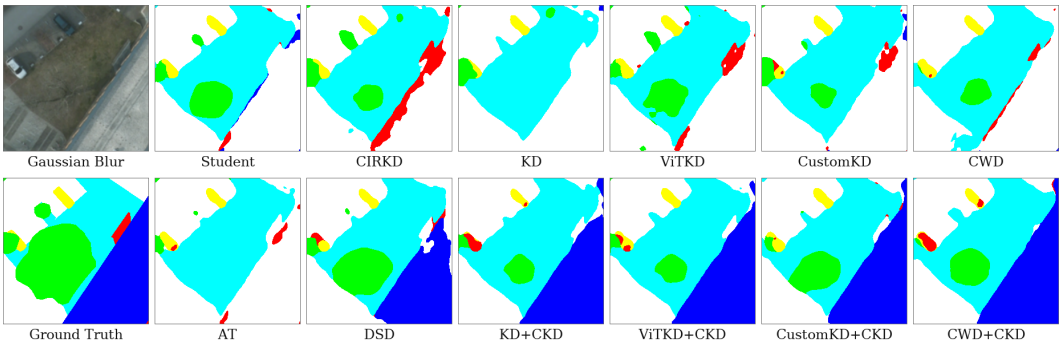


Figure S5: Qualitative comparison under the domain shift induced by Gaussian Blur.

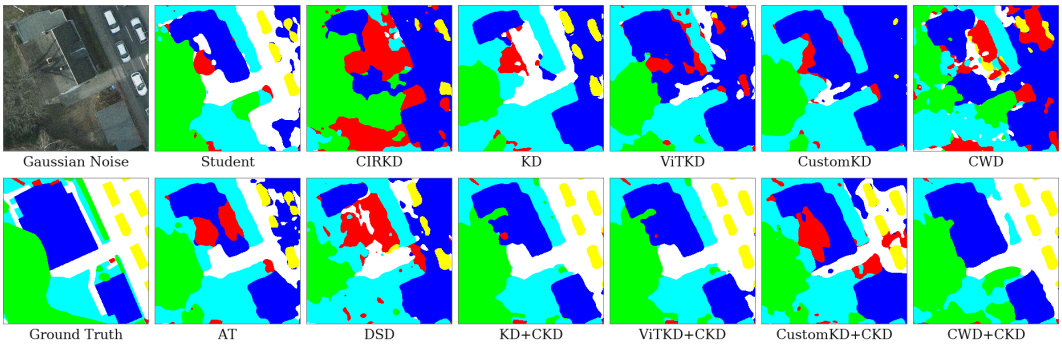


Figure S6: Qualitative comparison under the domain shift induced by Gaussian Noise.

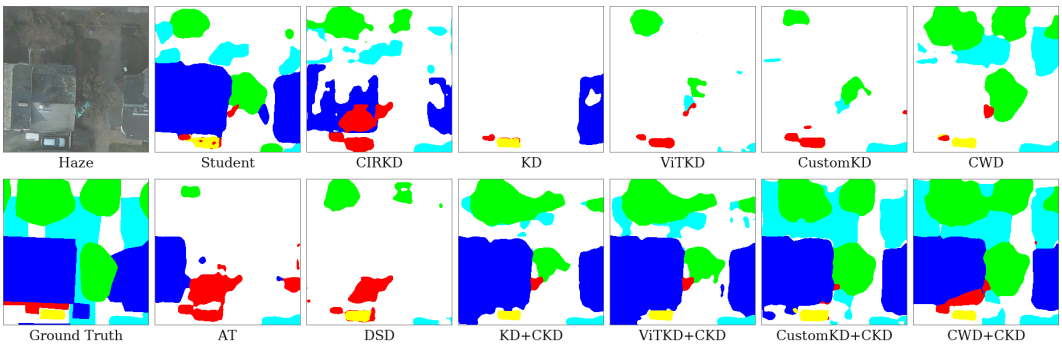


Figure S7: Qualitative comparison under the domain shift induced by Haze.

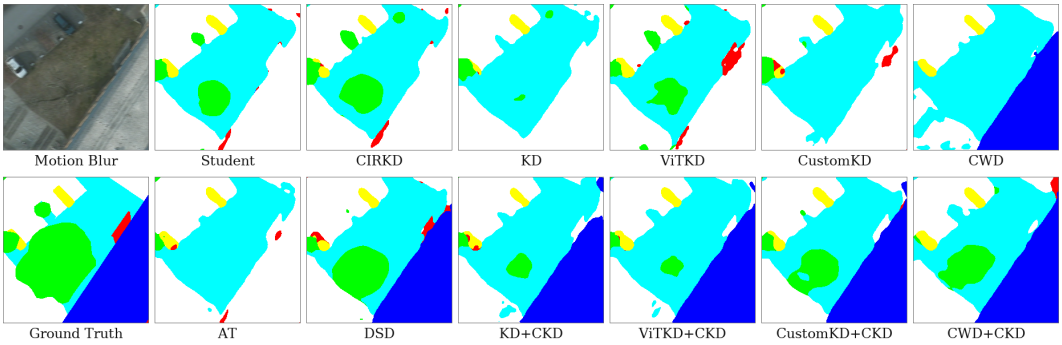


Figure S8: Qualitative comparison under the domain shift induced by Motion Blur.

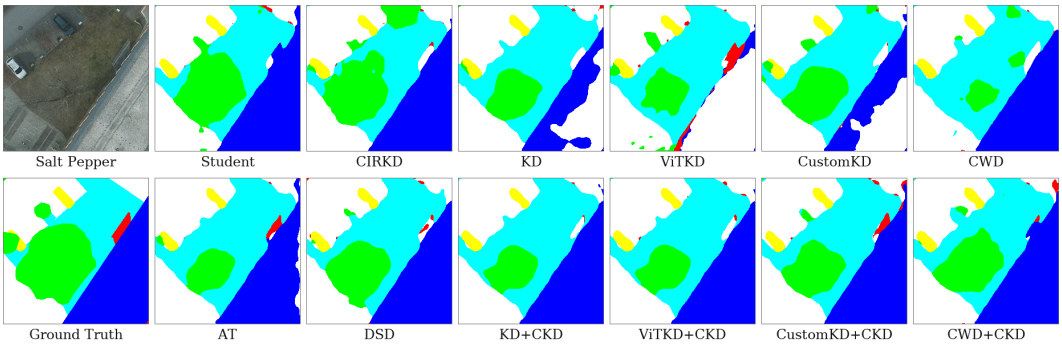


Figure S9: Qualitative comparison under the domain shift induced by Salt and Pepper Noise.

F Feature-Space Evidence of Stable Student Representations

Figure S10 examines whether CKD encourages the student to learn stable representations under domain shifts. For the same scene, we compare the student's final-block encoder features between the source domain and the corresponding domain-shift condition. Since the scene content and semantic labels remain unchanged, differences in feature representations mainly reflect sensitivity to domain-dependent appearance variations.

As shown in Figure S10, conventional KD and CWD exhibit noticeable changes in feature organization under Compression Artifacts and Haze, whereas their CKD-enhanced variants preserve spatial patterns that are more consistent with their source-domain counterparts. To complement this qualitative observation, we further compute the mean similarity between corresponding source-domain and shifted tokens. Under Compression Artifacts, the similarity increases from 0.553 to 0.708 for KD+CKD and from 0.613 to 0.804 for CWD+CKD. Under Haze, it increases from 0.637 to 0.774 and from 0.796 to 0.904, respectively. These results indicate that CKD reduces feature drift under domain shifts and encourages the student to preserve more stable task-relevant representations. This behavior is consistent with the goal of SCKA to reduce reliance on noncausal shortcuts, although the visualization itself does not establish causality.

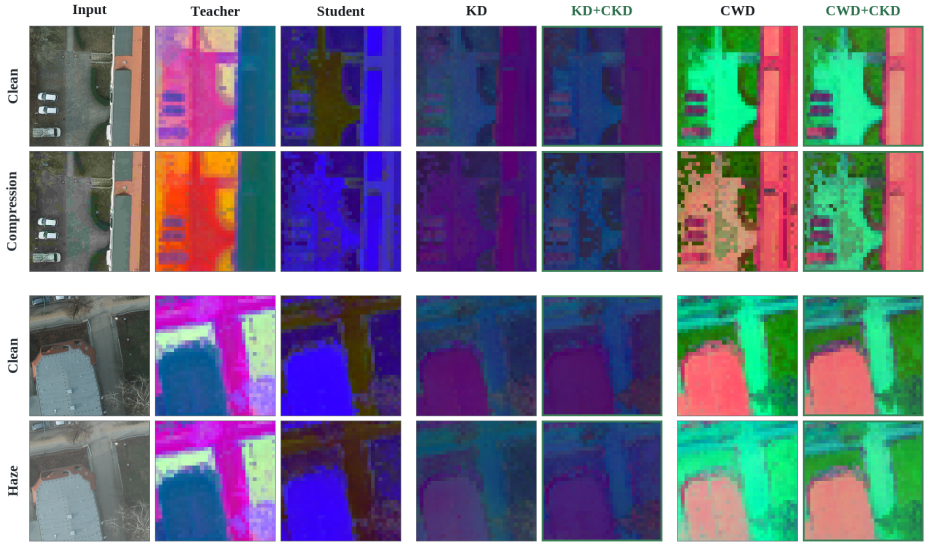


Figure S10: PCA visualization of feature representations under the source domain and matched domain-shift conditions induced by Compression Artifacts and Haze. Conventional KD and CWD exhibit larger changes in feature organization, whereas their CKD-enhanced variants retain spatial patterns more consistent with the source-domain representations. The ViT-L teacher is projected separately due to its different feature dimension.