RVHOMO: UNSUPERVISED CROSS-MODAL HOMOGRAPHY ESTIMATION WITH GEOMETRY-AWARE VFM FEATURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised cross-modal homography estimation is challenging because modality-dependent appearances obscure shared geometric structures, while common surrogate objectives do not explicitly enforce geometry-preserving behavior. We propose the Recurrent VFM-guided Homography Estimation Network (RVHomo), which combines broad structural cues from a frozen vision foundation model with the spatial precision of local CNN features. Rather than directly injecting coarse VFM tokens, RVHomo transforms hierarchical DINOv2 features into scale-aligned geometric representations and fuses them with CNN features for coarse-to-fine recurrent estimation. A dual-scale reconstruction objective further encourages the adapted representations to retain spatial structure. We additionally enforce geometric consistency at the modality-transfer and homography-prediction levels, imposing warp equivariance and composition-consistent predictions without cross-modal transformation labels. To reduce model complexity, we derive RVHomo-D through two-stage progressive distillation, eliminating modality transfer and online VFM extraction at inference. Experiments on GoogleMap, OPT-SAR, FSD, and VIS-IR demonstrate that RVHomo achieves MACE values of 0.65, 1.95, 4.32, and 4.55, reducing error by 17.4–46.7% relative to the strongest competing estimator on each dataset. RVHomo-D remains more accurate than all evaluated competitors while reducing RVHomo's total parameter count by 98.5% and computation by 70.4%.

1 INTRODUCTION

Homography estimation recovers a global projective transformation between two views for applications such as image registration, multi-sensor fusion, and map-based localization. Although deep homography estimators perform well under supervision, obtaining ground-truth transformations or precisely aligned image pairs is difficult across sensors, imaging mechanisms, and acquisition times. This motivates unsupervised cross-modal homography estimation, which must recover geometry from unaligned image pairs without cross-modal transformation labels.

Existing unsupervised methods derive surrogate supervision from image content using photometric reconstruction, learned feature similarity, perceptual consistency, warp-equivariant features, or motion-basis representations (Nguyen et al., 2018; Zhang et al., 2020; Koguciuk et al., 2021; Ye et al., 2021). Recent cross-modal approaches further introduce intra-modal self-supervision and consistent feature projection, alternate representation and geometry optimization, or reformulate the task through modality transfer and split optimization (Zhang et al., 2024; Song et al., 2024; Yu et al., 2025). Despite this progress, two coupled challenges remain. First, task-specific correspondence features can remain modality-dependent, yielding unreliable cues in repetitive or weakly textured regions. Second, reconstruction, feature-alignment, and decomposed objectives do not guarantee geometric preservation: modality transfer may alter spatial structures, while predictions may respond inconsistently to known perturbations. Addressing either challenge alone is insufficient; reliable representations and geometrically consistent learning must be developed together.

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.

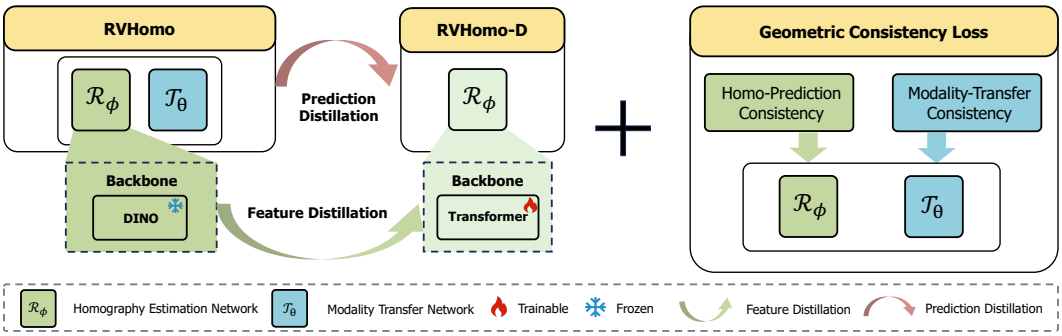


Figure 1: Overview of the proposed approach. RVHomo equips the SSHNet unsupervised framework with a DINOv2-guided homography estimator and geometric consistency at the modality-transfer and homography-prediction levels. Two-stage progressive distillation removes modality transfer and online VFM extraction to obtain the lightweight RVHomo-D.

Recent supervised advances suggest that stronger representations can benefit multimodal geometric estimation. GFNet exploits DINOv2 features for grid-based homography correspondence, while DistillMatch distills VFM semantics into a lightweight multimodal matcher (Zhang et al., 2025; Yang et al., 2025). Other work improves generalization to unseen modality combinations through style-diverse data synthesis and color-invariant representation learning (You et al., 2026). These results demonstrate the potential of VFMs, but their coarse semantic features are not inherently suited to corner-sensitive homography estimation. Moreover, existing approaches optimize their geometric models using explicit flow, correspondence, pose, or synthetic homography supervision. The central question is therefore how to adapt generic VFM representations into geometrically precise features using only real, unaligned cross-modal image pairs.

To address these challenges, we propose the Recurrent VFM-guided Homography Estimation Network (RVHomo), a geometry-aware VFM adaptation framework for unsupervised cross-modal homography estimation. Fig. 1 provides an overview of our framework, comprising the complete RVHomo model, geometric consistency learning at the modality-transfer and homography-prediction levels, and a two-stage progressive distillation scheme for deriving the lightweight RVHomo-D. Within the problem reformulation and split optimization of SSHNet (Yu et al., 2025), the complete RVHomo model employs a frozen DINOv2 encoder (Oquab et al., 2024) to complement local CNN descriptors with broader structural context. Our key insight is that VFM and CNN features provide complementary but inherently mismatched information: VFM representations capture broad structural cues that are relatively robust to appearance variations, whereas CNN features retain the spatial precision required for corner-sensitive localization. Rather than directly injecting coarse VFM tokens into the homography estimator, RVHomo transforms hierarchical DINOv2 features into scale-aligned geometric representations and integrates them with local CNN features for coarse-to-fine recurrent estimation. A dual-scale reconstruction objective further encourages the adapted VFM representations to retain recoverable spatial structure.

Stronger representations alone, however, do not guarantee geometrically valid unsupervised learning. We therefore introduce geometric consistency constraints at both the modality-transfer and homography-prediction levels. The former requires modality transfer to commute approximately with known geometric warping, discouraging the transfer network from altering spatial structures. The latter requires predictions from reference and perturbed views to satisfy the corresponding homography-composition relation. The sampled perturbations specify only these consistency relations and are not used as transformation labels between the original cross-modal images. Both constraints are applied only during training and introduce no additional inference-time computation.

The complete RVHomo model still requires a modality-transfer network and online VFM extraction at inference. As illustrated in Fig. 1, we further derive RVHomo-D through two-stage progressive distillation. The complete teacher first transfers prediction supervision to an intermediate model that directly processes the original cross-modal pair without modality transfer. The intermediate model then provides prediction and branch-wise feature supervision to RVHomo-D, whose lightweight Transformer proxy replaces the frozen VFM. This staged design avoids aligning teacher and student

features across different input domains, and RVHomo-D consequently requires neither modality transfer nor online VFM extraction.

Experiments on GoogleMap, OPT-SAR, FSD, and VIS-IR demonstrate that RVHomo consistently achieves the best performance, obtaining MACE values of 0.65, 1.95, 4.32, and 4.55, respectively. Compared with the strongest competing estimator on each dataset, RVHomo reduces MACE by 17.4–46.7%. RVHomo-D remains more accurate than all evaluated competitors while reducing the complete teacher's total parameter count by 98.5% and computation by 70.4%.

Our contributions are summarized as follows:

- We propose RVHomo, which transforms hierarchical features from a frozen VFM into scale-aligned geometric representations and integrates them with local CNN features for unsupervised cross-modal homography estimation.
- We introduce geometric consistency learning at the modality-transfer and homography-prediction levels, enforcing warp equivariance and composition-consistent predictions without requiring cross-modal transformation labels or adding inference-time computation.
- We develop two-stage progressive distillation that transfers prediction and branch-wise feature knowledge through an intermediate model, producing RVHomo-D without modality transfer or online VFM extraction.

2 Related Work

Deep homography estimators regress four-corner displacements through hierarchical or recurrent refinement (DeTone et al., 2016; Le et al., 2020; Cao et al., 2022; 2023; Zhu et al., 2024). Unsupervised variants replace transformation labels with image-derived objectives and content, motion, adversarial, or attention priors (Nguyen et al., 2018; Zhang et al., 2020; Koguciuk et al., 2021; Ye et al., 2021; Hong et al., 2022; Liu & Li, 2023). For cross-modal images, MIND, DASC, and RIFT seek modality-insensitive structure, while CoMIR learns a shared representation (Heinrich et al., 2012; Kim et al., 2015; Li et al., 2020; Pielawski et al., 2020). Recent methods use geometry-preserving translation, intra-modal self-supervision, alternating representation–geometry optimization, or problem reformulation with split optimization (Arar et al., 2020; Zhang et al., 2024; Song et al., 2024; Yu et al., 2025). RVHomo builds on SSHNet's reformulation but addresses modality-sensitive correspondence features and the absence of explicit geometric equivariance.

Recent work also exploits VFM priors and diverse supervision for multimodal geometry. GFNet initializes grid-flow prediction with DINOv2 features, DistillMatch distills VFM semantics into a lightweight matcher, and You et al. synthesize style-diverse pairs with known homographies (Zhang et al., 2025; Yang et al., 2025; You et al., 2026). These approaches demonstrate the benefits of stronger representations and appearance diversity, but train their geometric models with explicit flow, correspondence and pose, or synthetic homography supervision. RVHomo instead adapts hierarchical frozen-VFM features from real unaligned cross-modal pairs without ground-truth homography supervision between them, while constraining learning through geometric consistency.

3 Method

3.1 Overview

As illustrated in Fig. 2, we retain the problem reformulation and split optimization of SSHNet (Yu et al., 2025), while instantiating its homography estimator with RVHomo. Given a source image $\mathbf{I}_{\mathrm{A}}$ from modality A and a target image $\mathbf{I}_{\mathrm{B}}$ from modality B, RVHomo combines local CNN features with scale-aligned, multi-level VFM features and estimates four-corner displacements through coarse-to-fine recurrent refinement. During training, a dual-scale reconstruction loss regularizes the aggregated VFM representations, while two geometric consistency terms constrain modality transfer and homography prediction under known source-side perturbations.

The complete model still requires a modality transfer network and online VFM extraction at inference. We remove these dependencies through two-stage progressive distillation, as detailed later in this section.

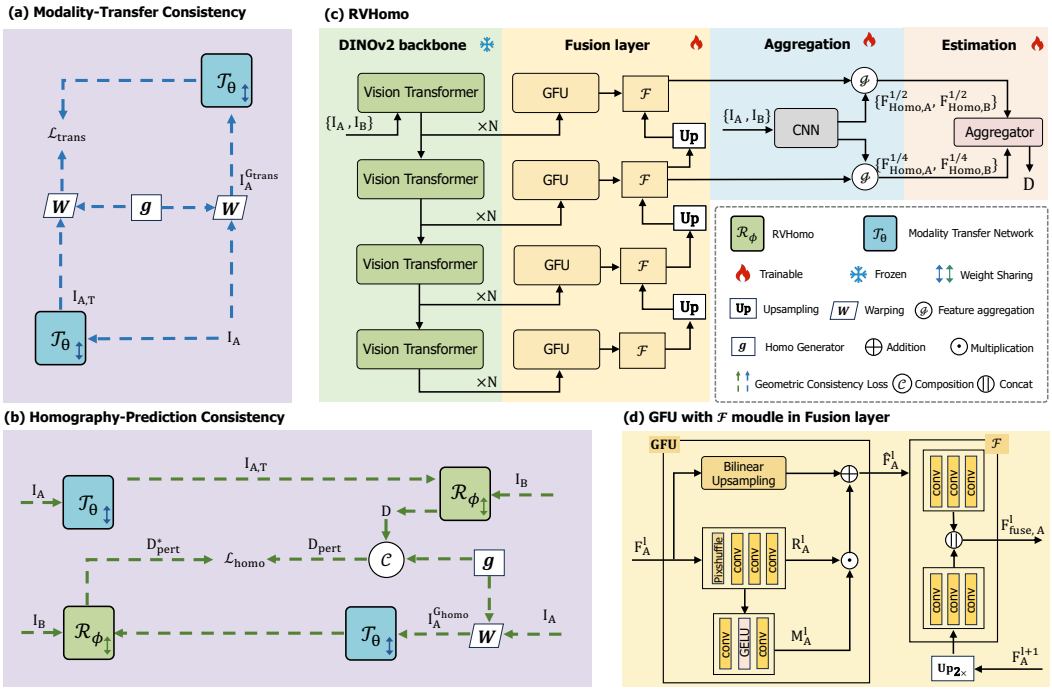


Figure 2: Overview of RVHomo. (a) The RVHomo framework equipped with the Modality-Transfer consistency objective. (b) The RVHomo framework equipped with the Homography-Prediction consistency objective. (c) RVHomo combines frozen DINOv2 features with local CNN features for multi-scale homography estimation. (d) GFU and multi-level fusion transform hierarchical VFM features into scale-aligned geometric representations.

3.2 VFM-Guided Architecture with Regularization

As detailed in Fig. 2, a CNN–VFM dual-branch encoder is shared between the source and target images and produces matching features at scales $s \in \{\frac{1}{4}, \frac{1}{2}\}$. For clarity, we describe the source branch for $\mathbf{I}_{\mathrm{A}}$; the target image $\mathbf{I}_{\mathrm{B}}$ follows the same operations with shared weights. The CNN branch extracts local image structures, while the frozen VFM partitions $\mathbf{I}_{\mathrm{A}}$ into non-overlapping patches and processes the resulting token sequence with a Transformer encoder. After removing non-spatial tokens, the outputs of L selected Transformer blocks, indexed by $\{\ell_j\}_{j=1}^{L}$, are linearly projected and spatially rearranged into feature maps $\{\mathbf{F}_{\mathrm{A}}^{\ell_j}\}_{j=1}^{L}$.

These low-resolution features are transformed to the resolutions used for homography matching by geometry-aware feature upsampling (GFU). Given $\mathbf{F}_{\mathrm{A}}^{\ell_j}$ from block ℓ_j and its upsampling factor r_j, GFU adds a gated geometry-refinement residual to a bilinearly upsampled feature:

$$\widehat{\mathbf{F}}_{\mathrm{A}}^{\ell_j} = \mathrm{GFU}(\mathbf{F}_{\mathrm{A}}^{\ell_j}, r_j) = \mathrm{BilUp}_{r_j}(\mathbf{F}_{\mathrm{A}}^{\ell_j}) + \mathbf{M}_{\mathrm{A}}^{\ell_j} \odot \mathbf{R}_{\mathrm{A}}^{\ell_j}. \tag{1}$$

The spatial gate $\mathbf{M}_{\mathrm{A}}^{\ell_j}$ controls where and how strongly the geometry-refinement residual $\mathbf{R}_{\mathrm{A}}^{\ell_j}$ is injected. The transformed features are then aggregated along a top-down pathway:

$$\mathbf{F}_{\mathrm{fuse,A}}^{\ell_j} = \mathcal{F}_j\left(\mathrm{Up}_2\left(\mathbf{F}_{\mathrm{fuse,A}}^{\ell_{j+1}}\right), \widehat{\mathbf{F}}_{\mathrm{A}}^{\ell_j}\right), \qquad j = L-1, \ldots, 1. \tag{2}$$

At each level, $\mathcal{F}_j$ transforms the two inputs separately, concatenates them along the channel dimension, and applies further refinement. Scale-specific channel projections yield the aggregated features $\mathbf{F}_{\mathrm{VFM,A}}^{s}$ at $1/4$ and $1/2$ resolution. Each feature is concatenated with its CNN counterpart and passed through a scale-specific fusion unit:

$$\mathbf{F}_{\mathrm{Homo,A}}^{s} = \mathcal{G}^{s}\left(\mathbf{F}_{\mathrm{CNN,A}}^{s} \,\|\, \mathbf{F}_{\mathrm{VFM,A}}^{s}\right), \qquad s \in \left\{\tfrac{1}{4}, \tfrac{1}{2}\right\}, \tag{3}$$

where $\|$ denotes channel-wise concatenation. Applying the same pipeline to $\mathbf{I}_{\mathrm{B}}$ yields $\mathbf{F}_{\mathrm{homo,B}}^{s}$. The fused source and target features are then processed by the recurrent homography estimation pipeline

of IHN (Cao et al., 2022), producing four-corner displacements in a coarse-to-fine manner. We leave its recurrent refinement and prediction head unchanged.

To regularize the aggregated VFM representations to retain recoverable image structure, we reconstruct each input at both matching scales. Let $\Psi_{4\times}$ and $\Psi_{2\times}$ denote the corresponding lightweight decoders built with sub-pixel upsampling (Shi et al., 2016). For the source image, the dual-scale reconstruction loss is

$$\begin{aligned}\mathcal{L}_{\text{recon,A}} = &\left\|\mathbf{I}_{\text{A}} - \Psi_{4\times}(\mathbf{F}^{1/4}_{\text{VFM,A}})\right\|_2^2 \\ &+ \left\|\mathbf{I}_{\text{A}} - \Psi_{2\times}(\mathbf{F}^{1/2}_{\text{VFM,A}})\right\|_2^2, \\ \mathcal{L}_{\text{recon}} = &\mathcal{L}_{\text{recon,A}} + \mathcal{L}_{\text{recon,B}},\end{aligned} \tag{4}$$

where $\mathcal{L}_{\text{recon,B}}$ is defined identically for $\mathbf{I}_{\text{B}}$.

The VFM encoder remains frozen throughout training, whereas GFU, multi-level aggregation, feature fusion, and the homography estimator are optimized jointly. The reconstruction decoders are used only for training and do not participate in homography prediction or inference.

3.3 Geometric Consistency Loss

The split optimization of SSHNet decouples modality transfer from homography estimation, but does not explicitly constrain how either process should respond to a known geometric perturbation. Related warp-consistency objectives use known transformations to supervise correspondence predictions without real-image ground truth (Truong et al., 2021). We independently sample two known source-side perturbations $\mathbf{G}_{\text{trans}}, \mathbf{G}_{\text{homo}} \in \mathbb{R}^{4\times 2}$ for modality-transfer consistency and homography-prediction consistency, respectively. Each perturbation is represented by displacements of the four source corners. For either perturbation $\mathbf{G}$, let $\mathbf{I}_{\text{A}}^{\mathbf{G}}$ denote the resulting source view and let $\mathcal{W}(\mathbf{X};\mathbf{G})$ apply the corresponding homography warp to $\mathbf{X}$. Each $\mathbf{G}$ is oriented from its perturbed source view to the reference source view.

Modality-Transfer Consistency. Let $\mathcal{T}_\theta$ denote the modality transfer network. We enforce image-level warp equivariance by comparing transfer after warping with warping after transfer. To exclude invalid pixels introduced by warping, we use the binary validity mask $\mathbf{V}_{\text{trans}}$ induced by $\mathbf{G}_{\text{trans}}$ and define

$$\begin{aligned}\mathcal{L}_{\text{trans}} = \frac{1}{\|\mathbf{V}_{\text{trans}}\|_1}\Big\|\mathbf{V}_{\text{trans}} \odot \Big[\mathcal{T}_\theta(\mathbf{I}_{\text{A}}^{\mathbf{G}_{\text{trans}}}) \\ - \text{sg}\Big(\mathcal{W}(\mathcal{T}_\theta(\mathbf{I}_{\text{A}});\mathbf{G}_{\text{trans}})\Big)\Big]\Big\|_1,\end{aligned} \tag{5}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. The warped reference transfer serves only as the target; consequently, $\mathcal{L}_{\text{trans}}$ updates $\mathcal{T}_\theta$ only through the perturbed-view branch.

Homography-Prediction Consistency. Let $\mathcal{R}_\phi$ denote RVHomo, whose final output is a four-corner displacement. The reference prediction, perturbed-view prediction, and composed pseudo target are

$$\begin{aligned}\mathbf{D}_{\text{ref}} &= \text{sg}[\mathcal{R}_\phi(\mathcal{T}_\theta(\mathbf{I}_{\text{A}}), \mathbf{I}_{\text{B}})], \\ \mathbf{D}_{\text{pert}} &= \mathcal{R}_\phi\Big(\text{sg}\Big[\mathcal{T}_\theta(\mathbf{I}_{\text{A}}^{\mathbf{G}_{\text{homo}}})\Big], \mathbf{I}_{\text{B}}\Big), \\ \mathbf{D}^\star_{\text{pert}} &= \mathcal{C}(\mathbf{D}_{\text{ref}}, \mathbf{G}_{\text{homo}}),\end{aligned} \tag{6}$$

where $\mathbf{D}_{\text{ref}}, \mathbf{D}_{\text{pert}}, \mathbf{D}^\star_{\text{pert}} \in \mathbb{R}^{4\times 2}$. The operator $\mathcal{C}(\cdot,\cdot)$ composes the reference prediction with the known source perturbation in homography-matrix space and returns the corresponding four-corner displacement; it is not an element-wise addition. Because $\mathbf{G}_{\text{homo}}$ maps the perturbed source view to the reference source view and the transformation represented by $\mathbf{D}_{\text{ref}}$ maps the reference source to the target, $\mathbf{D}^\star_{\text{pert}}$ represents the direct mapping from the perturbed source to the target. We then define

$$\mathcal{L}_{\text{homo}} = \text{mean}\big(\big|\mathbf{D}_{\text{pert}} - \mathbf{D}^\star_{\text{pert}}\big|\big), \tag{7}$$

where the mean is taken over valid samples and the eight coordinates of the four-corner displacement. The reference prediction $\mathbf{D}_{\text{ref}}$ and the modality-transferred perturbed input are stop-gradient,

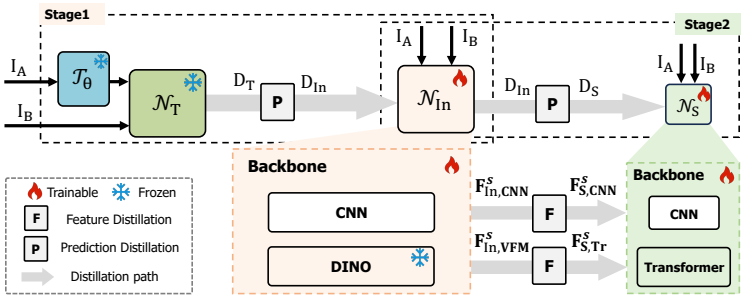


Figure 3: Two-stage progressive distillation. The complete teacher $\mathcal{N}_{\mathrm{T}}$ transfers prediction supervision to the intermediate model $\mathcal{N}_{\mathrm{In}}$, which provides prediction and branch-wise feature supervision to the lightweight student $\mathcal{N}_{\mathrm{S}}$.

so the composed target $\mathbf{D}^{\star}_{\mathrm{pert}}$ is also detached. Hence, $\mathcal{L}_{\mathrm{homo}}$ updates $\mathcal{R}_{\phi}$ only through the perturbed prediction $\mathbf{D}_{\mathrm{pert}}$.

Overall Objective. The two terms form the geometric consistency objective. Combining them with SSHNet's intra-modal homography, modality-transfer, and feature-space objectives and dual-scale reconstruction gives

$$\begin{aligned} \mathcal{L}_{\mathrm{consist}} &= \lambda_{\mathrm{trans}}\mathcal{L}_{\mathrm{trans}} + \lambda_{\mathrm{homo}}\mathcal{L}_{\mathrm{homo}}, \\ \mathcal{L}_{\mathrm{total}} &= \mathcal{L}_{\mathrm{H,A}} + \mathcal{L}_{\mathrm{H,B}} + \mathcal{L}_{\mathrm{T}} + \mathcal{L}_{\mathrm{HF}} + \lambda_{\mathrm{recon}}\mathcal{L}_{\mathrm{recon}} + \mathcal{L}_{\mathrm{consist}}. \end{aligned} \tag{8}$$

The first four terms are inherited from SSHNet (Yu et al., 2025): $\mathcal{L}_{\mathrm{H,A}}$ and $\mathcal{L}_{\mathrm{H,B}}$ supervise homography estimation on synthetically warped intra-modal pairs from modalities A and B, respectively; $\mathcal{L}_{\mathrm{T}}$ supervises modality transfer using the homography-aligned target; and $\mathcal{L}_{\mathrm{HF}}$ enforces homography feature-space consistency. We retain SSHNet's split optimization and stop-gradient paths without modification. The coefficients λ_{recon}, λ_{trans}, and λ_{homo} balance the three additional objectives. The added $\mathcal{L}_{\mathrm{trans}}$ and $\mathcal{L}_{\mathrm{homo}}$ update the modality-transfer network $\mathcal{T}_{\theta}$ and RVHomo $\mathcal{R}_{\phi}$, respectively. Both consistency paths are used only during training and introduce no inference-time computation.

3.4 Two-Stage Progressive Distillation

Knowledge distillation can transfer output behavior and intermediate representations from a larger model to a compact student (Hinton et al., 2015; Romero et al., 2015; Zagoruyko & Komodakis, 2017). As shown in Fig. 3, the complete teacher $\mathcal{N}_{\mathrm{T}}$ is connected to the lightweight RVHomo-D $\mathcal{N}_{\mathrm{S}}$ through an intermediate model $\mathcal{N}_{\mathrm{In}}$. To remove the modality-transfer branch, we directly adopt the prediction distillation strategy of SSHNet-D (Yu et al., 2025): the final prediction of $\mathcal{N}_{\mathrm{T}}$ serves as the pseudo target for $\mathcal{N}_{\mathrm{In}}$, which directly processes the original cross-modal pair. Within the same distillation framework, $\mathcal{N}_{\mathrm{In}}$ provides both prediction and feature supervision to $\mathcal{N}_{\mathrm{S}}$, enabling the VFM feature extractor to be replaced by a lightweight Transformer proxy.

For prediction distillation, let $\mathbf{D}_{\mathrm{In}} \in \mathbb{R}^{4\times 2}$ be the final four-corner prediction of $\mathcal{N}_{\mathrm{In}}$ and let $\{\mathbf{D}_{\mathrm{S}}^{(k)}\}_{k=1}^{K}$ be the K recurrent predictions of $\mathcal{N}_{\mathrm{S}}$. Feature distillation is applied before branch fusion by aligning the student CNN branch with the intermediate CNN branch and the student Transformer branch with the intermediate VFM branch.

For feature distillation, we describe modality A; modality B follows the same process. Let $\mathcal{S} = \{\frac{1}{4}, \frac{1}{2}\}$ denote the matching scales, $d(\cdot,\cdot)$ a normalized feature discrepancy for a corresponding feature pair, and ω_s the weight for scale s. We formulate the prediction distillation loss $\mathcal{L}_{\mathrm{pred}}$ and the modality-specific feature distillation loss $\mathcal{L}_{\mathrm{feat,A}}$, and combine both modalities into the overall

distillation objective $\mathcal{L}_{\text{distill}}$:

$$\begin{aligned}
\mathcal{L}_{\text{pred}} &= \sum_{k=1}^{K} \gamma^{K-k} \,\text{mean}\Big(\Big|\mathbf{D}_{\text{S}}^{(k)} - \text{sg}(\mathbf{D}_{\text{In}})\Big|\Big)\,, \\
\mathcal{L}_{\text{feat,A}} &= \sum_{s\in\mathcal{S}} \omega_s \Big[d\big(\mathbf{F}^s_{\text{S,CNN,A}}, \text{sg}(\mathbf{F}^s_{\text{In,CNN,A}})\big) \\
&\qquad\qquad + d\big(\mathbf{F}^s_{\text{S,Tr,A}}, \text{sg}(\mathbf{F}^s_{\text{In,VFM,A}})\big) \Big], \\
\mathcal{L}_{\text{distill}} &= \lambda_{\text{pred}}\mathcal{L}_{\text{pred}} + \lambda_{\text{feat}} \left(\mathcal{L}_{\text{feat,A}} + \mathcal{L}_{\text{feat,B}}\right).
\end{aligned} \tag{9}$$

Here, γ is the decay factor for recurrent predictions, and $\mathcal{L}_{\text{feat,B}}$ is defined identically using the features of modality B. Lightweight projection heads align the student and intermediate feature dimensions before feature distillation. At inference, the teacher, intermediate model, and all distillation-specific components are discarded; RVHomo-D requires neither the modality-transfer network nor VFM inference.

4 Experiments

4.1 Experimental Settings

Datasets and metric. Experiments are conducted on GoogleMap (Zhao et al., 2021), OPT-SAR (Li et al., 2022), FSD (Wen et al., 2025), and VIS-IR (Sun et al., 2022). Following prior work (Cao et al., 2022; 2023; Zhu et al., 2024), we evaluate homography estimation accuracy using mean average corner error (MACE), where a lower value indicates higher accuracy.

Baselines. We compare with generic unsupervised approaches, including **UDHN** (Nguyen et al., 2018), **CA-UDHN** (Zhang et al., 2020), **biHomE** (Koguciuk et al., 2021), **BasesHomo** (Ye et al., 2021), and **SCPNet** (Zhang et al., 2024), as well as **AltO** (Song et al., 2024) instantiated with **DHN** (DeTone et al., 2016), **RAFT** (Teed & Deng, 2020), **IHN** (Cao et al., 2022), and **RHWF** (Cao et al., 2023), and **SSHNet** (Yu et al., 2025) instantiated with **DHN**, **MHN** (Le et al., 2020), **IHN**, **RHWF**, and **MCNet** (Zhu et al., 2024).

4.2 Comparisons with Existing Methods

We compare the homography estimation results of RVHomo and RVHomo-D with existing methods in Table 1.

RVHomo achieves the best performance on all four datasets, with MACE values of 0.65, 1.95, 4.32, and 4.55 on GoogleMap, OPT-SAR, FSD, and VIS-IR, respectively. Compared with the strongest competing method on each dataset, namely SSHNet-MCNet on GoogleMap and OPT-SAR and SSHNet-IHN on FSD and VIS-IR, RVHomo reduces MACE by 46.7%, 32.1%, 18.8%, and 17.4%, respectively. The improvements demonstrate that the proposed representation is effective across diverse modality combinations.

Fig. 4 presents representative registration results on the four modality pairs, where RVHomo achieves the lowest sample ACE in every row and produces predictions that most closely overlap the ground-truth quadrilaterals. Its improved alignment along map boundaries, riverbanks, object contours, and weakly textured infrared structures is consistent with the quantitative results.

4.3 Ablation Studies

The ablation studies are conducted on the OPT-SAR dataset.

Geometric consistency loss. We incrementally add modality-transfer consistency $\mathcal{L}_{\text{trans}}$ and homography-prediction consistency $\mathcal{L}_{\text{homo}}$ to RVHomo while keeping the architecture and reconstruction regularization fixed.

The results show that $\mathcal{L}_{\text{trans}}$ improves registration accuracy and that $\mathcal{L}_{\text{homo}}$ provides a further gain, demonstrating that the two consistency terms are complementary.

Table 1: Cross-modal homography estimation in MACE. GM, OS, and VI denote GoogleMap, OPT-SAR, and VIS-IR; NC indicates non-convergence. All reported SSHNet variants are trained without our proposed consistency strategy.

Method family	Method	Publication	GM	OS	FSD	VI
Generic unsupervised	UDHN (Nguyen et al., 2018)	RA-L 2018	22.84	NC	NC	NC
	CA-UDHN (Zhang et al., 2020)	ECCV 2020	24.61	24.76	24.86	24.55
	biHomE (Koguciuk et al., 2021)	CVPR 2021	NC	NC	NC	NC
	BasesHomo (Ye et al., 2021)	ICCV 2021	24.49	26.72	25.08	NC
	SCPNet (Zhang et al., 2024)	ECCV 2024	4.35	NC	NC	NC
AltO variants (Song et al., 2024)	DHN (DeTone et al., 2016)	arXiv 2016	6.19	24.83	NC	12.24
	RAFT (Teed & Deng, 2020)	ECCV 2020	3.10	11.38	NC	7.84
	IHN (Cao et al., 2022)	CVPR 2022	3.06	11.69	NC	6.83
	RHWF (Cao et al., 2023)	CVPR 2023	3.49	13.97	NC	6.17
SSHNet variants (Yu et al., 2025)	DHN (DeTone et al., 2016)	arXiv 2016	9.28	17.63	15.02	10.86
	MHN (Le et al., 2020)	CVPR 2020	2.90	6.77	10.37	7.51
	IHN (Cao et al., 2022)	CVPR 2022	1.23	2.94	5.32	5.51
	RHWF (Cao et al., 2023)	CVPR 2023	1.29	3.08	5.64	5.60
	MCNet (Zhu et al., 2024)	CVPR 2024	1.22	2.87	6.06	6.22
Ours	**RVHomo**	–	**0.65**	**1.95**	**4.32**	**4.55**
	RVHomo-D	–	0.71	2.04	4.77	4.72

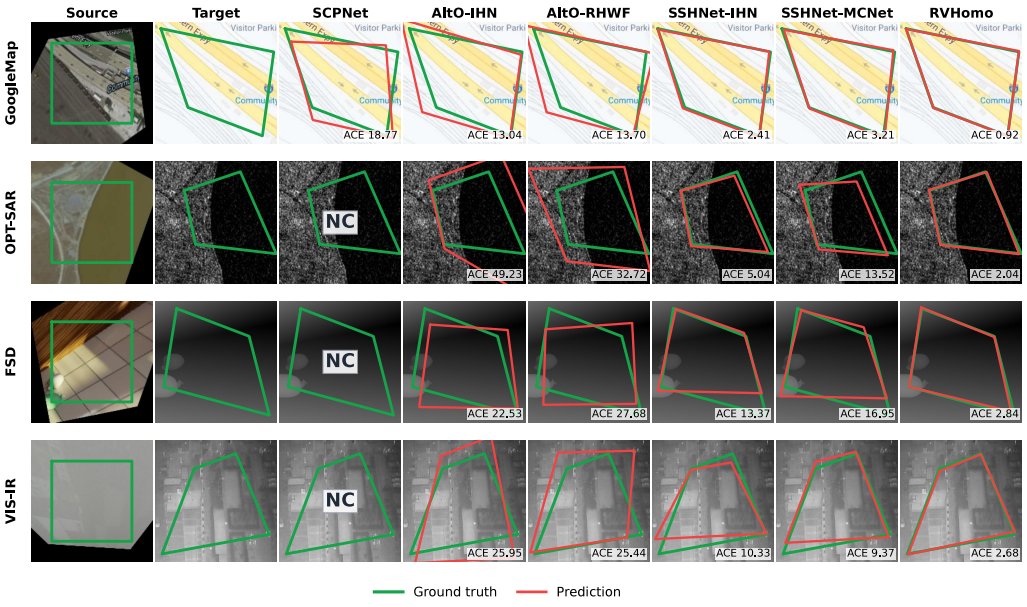


Figure 4: Qualitative comparison on four modality pairs. Each row uses the same input pair and perturbation for all methods.

Architecture ablation. With the correlation construction and recurrent estimator fixed, we isolate how VFM features are incorporated. **w/o VFM** removes the frozen-VFM path and forms the $1/4$- and $1/2$-resolution matching features from the CNN pyramid. **w/o GFU** retains the four-level hierarchy and top-down aggregation but replaces learnable scale transformation with bilinear resizing and convolutional projection. **w/o top-down** retains all GFU-adapted levels but replaces sequential top-down aggregation with parallel resizing and concatenation. **RVHomo** uses the complete adaptation and reconstruction design.

The w/o VFM variant obtains 2.54 MACE, compared with 1.95 for RVHomo, demonstrating the benefit of incorporating adapted VFM representations within the same estimator. Removing GFU

Table 2: Ablation study of $\mathcal{L}_{\text{trans}}$ and $\mathcal{L}_{\text{homo}}$.

$\mathcal{L}_{\text{trans}}$	$\mathcal{L}_{\text{homo}}$	MACE
×	×	2.38
×	✓	2.19
✓	×	2.04
✓	✓	1.95

Table 3: Architecture ablation.

Configuration	MACE
w/o VFM	2.54
w/o GFU	3.74
w/o top-down	NC
RVHomo	**1.95**

Table 4: Direct versus progressive distillation.

Intermediate	$\mathcal{L}_{\text{pred}}$	$\mathcal{L}_{\text{feat}}$	MACE
×	✓	×	2.80
×	✓	✓	5.39
✓	✓	×	2.31
✓	✓	✓	**2.04**

increases the final error to 3.74, while the model without top-down aggregation does not converge under the common training protocol. These results support the roles of learnable scale adaptation and hierarchical aggregation in stable convergence.

4.4 Two-Stage Progressive Distillation

Table 5: Model complexity. Trainable parameters exclude frozen modules.

Model	Modality transfer	VFM inference	Trainable params.	Total params.	GFLOPs
RVHomo	✓	✓	16.179M	320.547M	120.252
Intermediate	×	✓	8.604M	312.973M	116.671
RVHomo-D	×	×	**4.877M**	**4.877M**	**35.552**

We compare the teacher model RVHomo, the intermediate model, and the student model RVHomo-D on OPT-SAR.

Distillation ablation. Direct prediction distillation from RVHomo to RVHomo-D reaches 2.80 MACE, whereas adding feature supervision degrades it to 5.39. The teacher processes a pair produced by modality transfer, while the student receives the original cross-modal pair. Their geometric predictions remain directly comparable, but their branch features are induced from different input domains and are therefore unsuitable for direct alignment. Progressive distillation resolves this mismatch. Prediction distillation first removes modality transfer while retaining the VFM, yielding the intermediate model at 2.31 MACE. The intermediate model then provides prediction and feature supervision to RVHomo-D in the same cross-modal input domain, reducing MACE to 2.04. Thus, the intermediate stage is specifically required to make feature distillation effective.

Model complexity. Removing modality transfer cuts trainable parameters by 46.8%, but leaves total parameters and GFLOPs nearly unchanged because VFM inference remains. RVHomo-D removes this bottleneck, requiring 4.877M total parameters and 35.552 GFLOPs, reductions of 98.5% and 70.4% relative to RVHomo.

5 Conclusion

We presented RVHomo, a VFM-guided framework for unsupervised cross-modal homography estimation. RVHomo effectively combines VFM adaptation with geometric consistency constraints and yields a lightweight student through progressive distillation. Extensive experiments on four cross-modal benchmarks demonstrate state-of-the-art performance. Future work will explore the applicability of RVHomo to real-world registration scenarios.

AI USE STATEMENT

Generative AI tools were used to check code correctness, improve manuscript readability. They were not used to originate the core methodology or formulate the central research hypotheses. All AI-assisted suggestions, code changes, and written content included in the paper were manually examined by the authors. The authors made all final decisions concerning the methodology, experimental design, interpretation, and claims, and take full responsibility for the final manuscript, code, and other AI-assisted artifacts.

REFERENCES

Moab Arar, Yiftach Ginger, Dov Danon, Amit H. Bermano, and Daniel Cohen-Or. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13410–13419, 2020.

Si-Yuan Cao, Jianxin Hu, Zehua Sheng, and Hui-Liang Shen. Iterative deep homography estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1879–1888, 2022.

Si-Yuan Cao, Runmin Zhang, Lun Luo, Beinan Yu, Zehua Sheng, Junwei Li, and Hui-Liang Shen. Recurrent homography estimation using homography-guided image warping and focus transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9833–9842, 2023.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016.

Mattias P. Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V. Gleeson, Michael Brady, and Julia A. Schnabel. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical Image Analysis*, 16(7):1423–1435, 2012. doi: 10.1016/j.media.2012.05.008.

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

Mingbo Hong, Yuhang Lu, Nianjin Ye, Chunyu Lin, Qijun Zhao, and Shuaicheng Liu. Unsupervised homography estimation with coplanarity-aware GAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17663–17672, 2022.

Seungryong Kim, Dongbo Min, Bumsub Ham, Seungchul Ryu, Minh N. Do, and Kwanghoon Sohn. DASC: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2103–2112, 2015.

Daniel Koguciuk, Elahe Arani, and Bahram Zonooz. Perceptual loss for robust unsupervised homography estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4274–4283, 2021.

Hoang Le, Feng Liu, Shu Zhang, and Aseem Agarwala. Deep homography estimation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7652–7661, 2020.

Jiayuan Li, Qingwu Hu, and Mingyao Ai. RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Transactions on Image Processing*, 29:3296–3310, 2020. doi: 10.1109/TIP.2019.2959244.

Xue Li, Guo Zhang, Hao Cui, Shasha Hou, Shunyao Wang, Xin Li, Yujia Chen, Zhijiang Li, and Li Zhang. MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification. *International Journal of Applied Earth Observation and Geoinformation*, 106:102638, 2022. doi: 10.1016/j.jag.2021.102638.

Jiazhen Liu and Xirong Li. Geometrized transformer for self-supervised homography estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9556–9565, 2023.

Ty Nguyen, Steven W. Chen, Shreyas S. Shivakumar, Camillo J. Taylor, and Vijay Kumar. Unsupervised deep homography: A fast and robust homography estimation model. *IEEE Robotics and Automation Letters*, 3(3):2346–2353, 2018. doi: 10.1109/LRA.2018.2809549.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Nicolas Pielawski, Elisabeth Wetzer, Johan Öfverstedt, Jiahao Lu, Carolina Wåhlby, Joakim Lindblad, and Nataša Sladoje. CoMIR: Contrastive multimodal image representation for registration. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18433–18444, 2020.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1874–1883, 2016.

Sanghyeob Song, Jaihyun Lew, Hyemi Jang, and Sungroh Yoon. Unsupervised homography estimation on multimodal image pair via alternating optimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 61306–61327, 2024.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022. doi: 10.1109/TCSVT.2022.3168279.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pp. 402–419, 2020.

Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Warp consistency for unsupervised learning of dense correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10346–10356, 2021.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. FoundationStereo: Zero-shot stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5249–5260, 2025.

Meng Yang, Fan Fan, Zizhuo Li, Songchu Deng, Yong Ma, and Jiayi Ma. DistillMatch: Leveraging knowledge distillation from vision foundation model for multimodal image matching. *arXiv preprint arXiv:2509.16017*, 2025.

Nianjin Ye, Chuan Wang, Haoqiang Fan, and Shuaicheng Liu. Motion basis learning for unsupervised deep homography estimation with subspace projection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13117–13125, 2021.

Jinkun You, Jiaxin Cheng, Jie Zhang, and Yicong Zhou. Towards generalized multimodal homography estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8408–8417, 2026.

Junchen Yu, Si-Yuan Cao, Runmin Zhang, Chenghao Zhang, Zhu Yu, Shujie Chen, Bailin Yang, and Hui-Liang Shen. SSHNet: Unsupervised cross-modal homography estimation via problem reformulation and split optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16685–16694, 2025.

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

Jirong Zhang, Chuan Wang, Shuaicheng Liu, Lanpeng Jia, Nianjin Ye, Jue Wang, Ji Zhou, and Jian Sun. Content-aware unsupervised deep homography estimation. In *Proceedings of the European Conference on Computer Vision*, pp. 653–669, 2020.

Kaining Zhang, Yuxin Deng, Jiayi Ma, and Paolo Favaro. Adapting dense matching for homography estimation with grid-based acceleration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6294–6303, 2025.

Runmin Zhang, Jun Ma, Si-Yuan Cao, Lun Luo, Beinan Yu, Shu-Jie Chen, Junwei Li, and Hui-Liang Shen. SCPNet: Unsupervised cross-modal homography estimation via intra-modal self-supervised learning. In *Proceedings of the European Conference on Computer Vision*, pp. 460–477, 2024.

Yiming Zhao, Xinming Huang, and Ziming Zhang. Deep lucas-kanade homography for multimodal image alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15950–15959, 2021.

Haokai Zhu, Si-Yuan Cao, Jianxin Hu, Sitong Zuo, Beinan Yu, Jiacheng Ying, Junwei Li, and Hui-Liang Shen. MCNet: Rethinking the core ingredients for accurate and efficient homography estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25932–25941, 2024.

APPENDIX

A Additional Details of the Geometric Consistency Loss

This section records the perturbation sampling and homography composition used by the two consistency terms in Section 3.3 of the main paper. Both perturbations are oriented from the perturbed source to the reference source; the target image remains unchanged.

A.1 Sampling of the Two Source-Side Perturbations

The two consistency paths use independently sampled source-side perturbations. For corner $i \in \{1, 2, 3, 4\}$, the horizontal and vertical components are sampled independently from discrete uniform distributions:

$$\begin{aligned} g_{\mathrm{trans},i,x}, g_{\mathrm{trans},i,y} &\sim \mathcal{U}_{\mathbb{Z}}[-32, 32], \\ g_{\mathrm{homo},i,x}, g_{\mathrm{homo},i,y} &\sim \mathcal{U}_{\mathbb{Z}}[-8, 8]. \end{aligned} \tag{A.1}$$

The loss weights are $\lambda_{\mathrm{trans}} = 1.0$ and $\lambda_{\mathrm{homo}} = 0.25$. A 128×128 patch is embedded in a 192×192 canvas with a 32-pixel context margin; invalid homography solves are replaced by a fixed valid fallback.

Let $\Pi(\mathbf{D})$ convert a four-corner displacement to a homography and Π^{-1} convert it back. The composed target is

$$\mathcal{C}(\mathbf{D}_{\mathrm{ref}}, \mathbf{G}_{\mathrm{homo}}) = \Pi^{-1}[\Pi(\mathbf{D}_{\mathrm{ref}})\Pi(\mathbf{G}_{\mathrm{homo}})]\,. \tag{A.2}$$

The matrix order follows the perturbed-to-reference-to-target chain and must not be replaced by displacement addition. A sample is used only when the reference prediction, perturbed prediction, and composed target are finite and the composed displacement has maximum absolute component at most 64 pixels.

B Additional Details of Two-Stage Progressive Distillation

The teacher–intermediate–student sequence and distillation objectives are defined in Section 3.4 of the main paper. RVHomo-D replaces the frozen VFM with a compact encoder–decoder whose feature resolutions and channels are $64 \times 64/64$, $32 \times 32/96$, and $16 \times 16/192$. Windowed Transformer blocks use 8×8 windows, an MLP ratio of 4, four attention heads at the two larger scales, and eight at the bottleneck. The decoder returns 32×32 and 64×64 features, which are fused with the corresponding CNN features before the unchanged homography estimator.

C Visual Analysis of the Geometric Consistency Loss

We complement the main-paper ablation with GoogleMap visualizations of spatial preservation and prediction equivariance. Ground-truth homographies are used only for evaluation.

C.1 Modality-Transfer Consistency

Figure C.1 compares translation with and without $\mathcal{L}_{\mathrm{trans}}$ at full-image and local-topology scales; cyan denotes the reference topology and orange the translated topology.

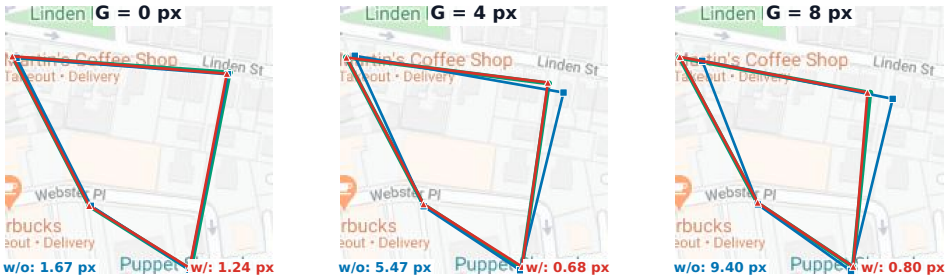
With $\mathcal{L}_{\mathrm{trans}}$, the sample's mean local residual decreases from 2.24 to 0.48 pixels. Over 160 samples (Table C.1), path-wise and edge discrepancies decrease from 0.0188 to 0.0134 and from 0.2693 to 0.2403, respectively.



(a) Source image (b) Translation w/o $\mathcal{L}_{\text{trans}}$ (c) Translation w/ $\mathcal{L}_{\text{trans}}$

(d) Source topology (e) Translation w/o $\mathcal{L}_{\text{trans}}$ (f) Translation w/ $\mathcal{L}_{\text{trans}}$

Figure C.1: Spatial preservation during modality transfer. The first row shows the source image and modality-transferred outputs. The second row shows an enlarged local region from the same sample. Cyan dashed lines denote the reference topology, and orange solid lines denote the topology observed after modality transfer. Modality-transfer consistency substantially reduces the local nonuniform deformation.



(a) $G = 0$ pixels (b) $G = 4$ pixels (c) $G = 8$ pixels

Figure C.2: Effect of homography-prediction consistency under increasing source-view perturbations. Blue circles denote the evaluation-only ground truth, red squares denote predictions without $\mathcal{L}_{\text{homo}}$, and green triangles denote predictions with $\mathcal{L}_{\text{homo}}$. The values in each panel are the corresponding sample ACEs.

Table C.1: Aggregate evaluation of modality-transfer consistency over 160 GoogleMap test samples. Lower discrepancy is better.

Evaluation measure	w/o $\mathcal{L}_{\text{trans}}$	w/ $\mathcal{L}_{\text{trans}}$	Mean reduction	Improved samples
Path-wise image discrepancy	0.0188	**0.0134**	28.9%	99.4%
Structural edge discrepancy	0.2693	**0.2403**	10.0%	95.0%

C.2 Homography-Prediction Consistency

Figure C.2 compares predictions with and without $\mathcal{L}_{\text{homo}}$ under fixed source-view perturbations of 0, 4, and 8 pixels. All overlays use the original 192×192 coordinates without visual amplification.

Without $\mathcal{L}_{\text{homo}}$, sample ACE grows from 1.67 to 9.40; with $\mathcal{L}_{\text{homo}}$, it remains between 0.68 and 1.24. Over 888 samples (Table C.2), eight-pixel degradation decreases from 0.817 to 0.261.

Table C.2: Prediction robustness under controlled source-view perturbations over 888 GoogleMap test samples. Lower MACE and degradation are better.

Perturbation (pixel)	MACE w/o $\mathcal{L}_{\text{homo}}$	MACE w/ $\mathcal{L}_{\text{homo}}$	Degradation w/o $\mathcal{L}_{\text{homo}}$	Degradation w/ $\mathcal{L}_{\text{homo}}$
0	1.535	**0.705**	0.000	0.000
2	1.759	**0.768**	0.224	0.063
4	1.976	**0.811**	0.441	0.106
6	2.134	**0.892**	0.599	0.187
8	2.352	**0.967**	0.817	0.261

Together, the two analyses support the complementary effects reported in the main paper.

D Dataset Examples

Figures D.1– D.4 provide representative input pairs from the four datasets. For each dataset, we uniformly sample 16 pairs from the evaluation set and visualize the actual 128×128 cross-modal inputs after sample construction. No registration result or prediction error is used for sample selection.

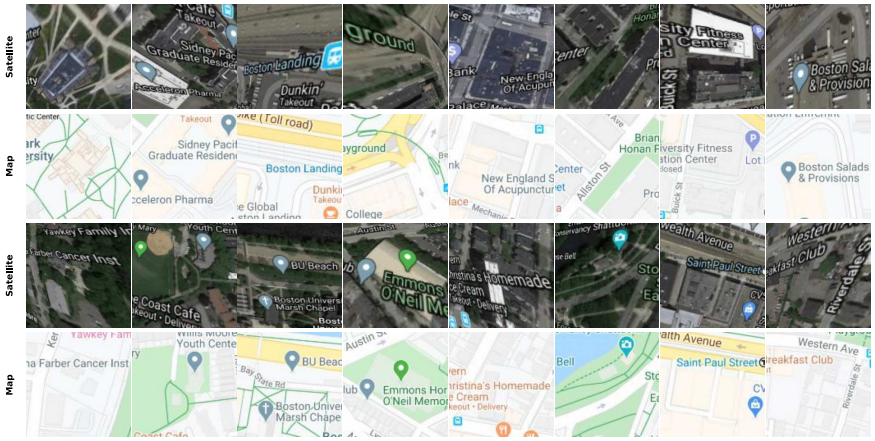


Figure D.1: Example input pairs from GoogleMap. Rows alternate between satellite and map images.

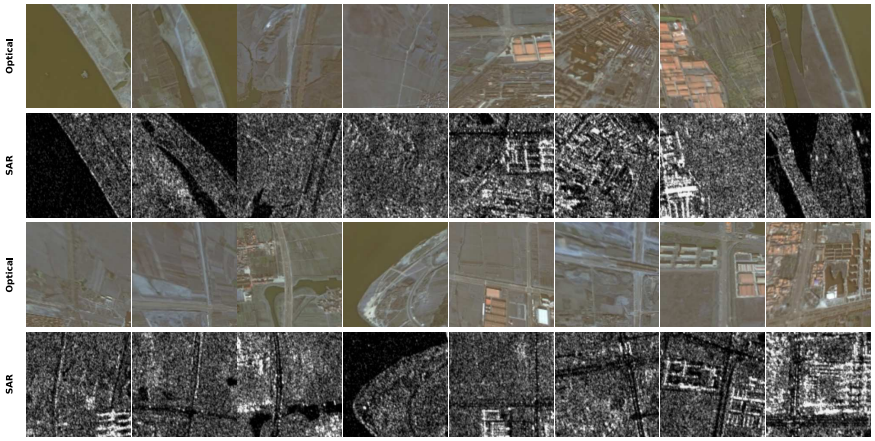


Figure D.2: Example input pairs from OPT-SAR. Rows alternate between optical and SAR images.

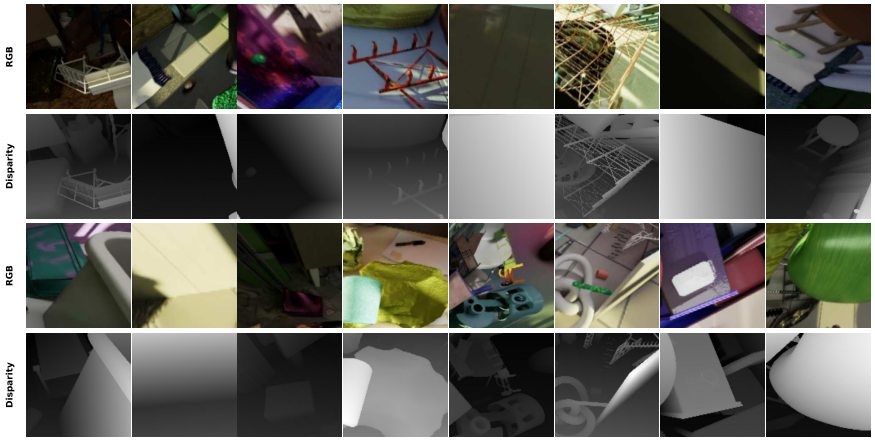


Figure D.3: Example input pairs from FSD. Rows alternate between RGB and disparity images.



Figure D.4: Example input pairs from VIS-IR. Rows alternate between visible and infrared images.

E Additional Qualitative Results

Figures E.1–E.4 present qualitative comparisons of different methods across the four datasets. The displayed examples illustrate cases in which RVHomo performs favorably. The first two columns show the source and target images. Green and red quadrilaterals denote the ground truth and prediction, respectively, and each reported value is ACE in pixels. "NC" indicates dataset-level non-convergence under the common protocol.

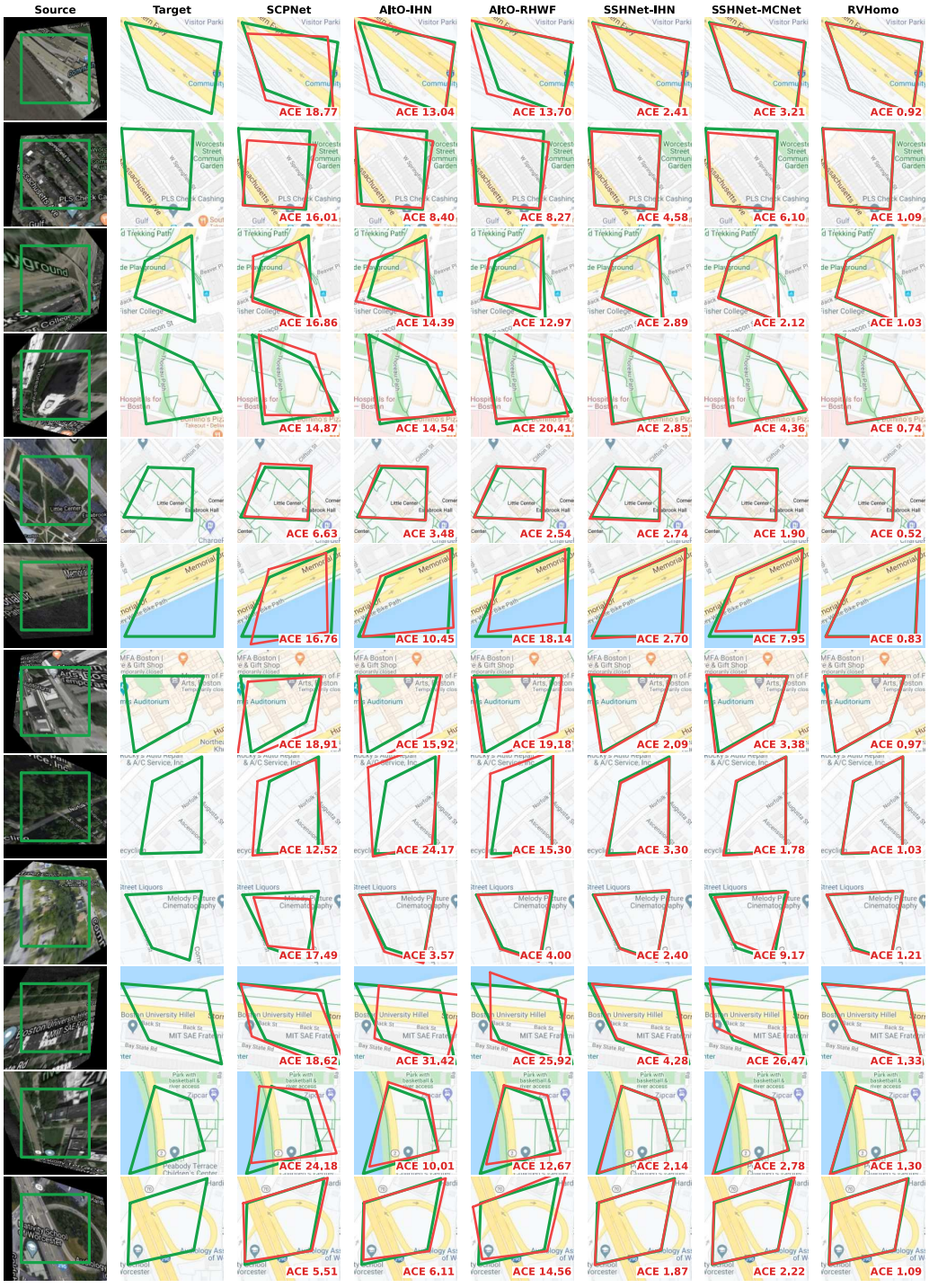


Figure E.1: Qualitative comparison of different methods on GoogleMap.

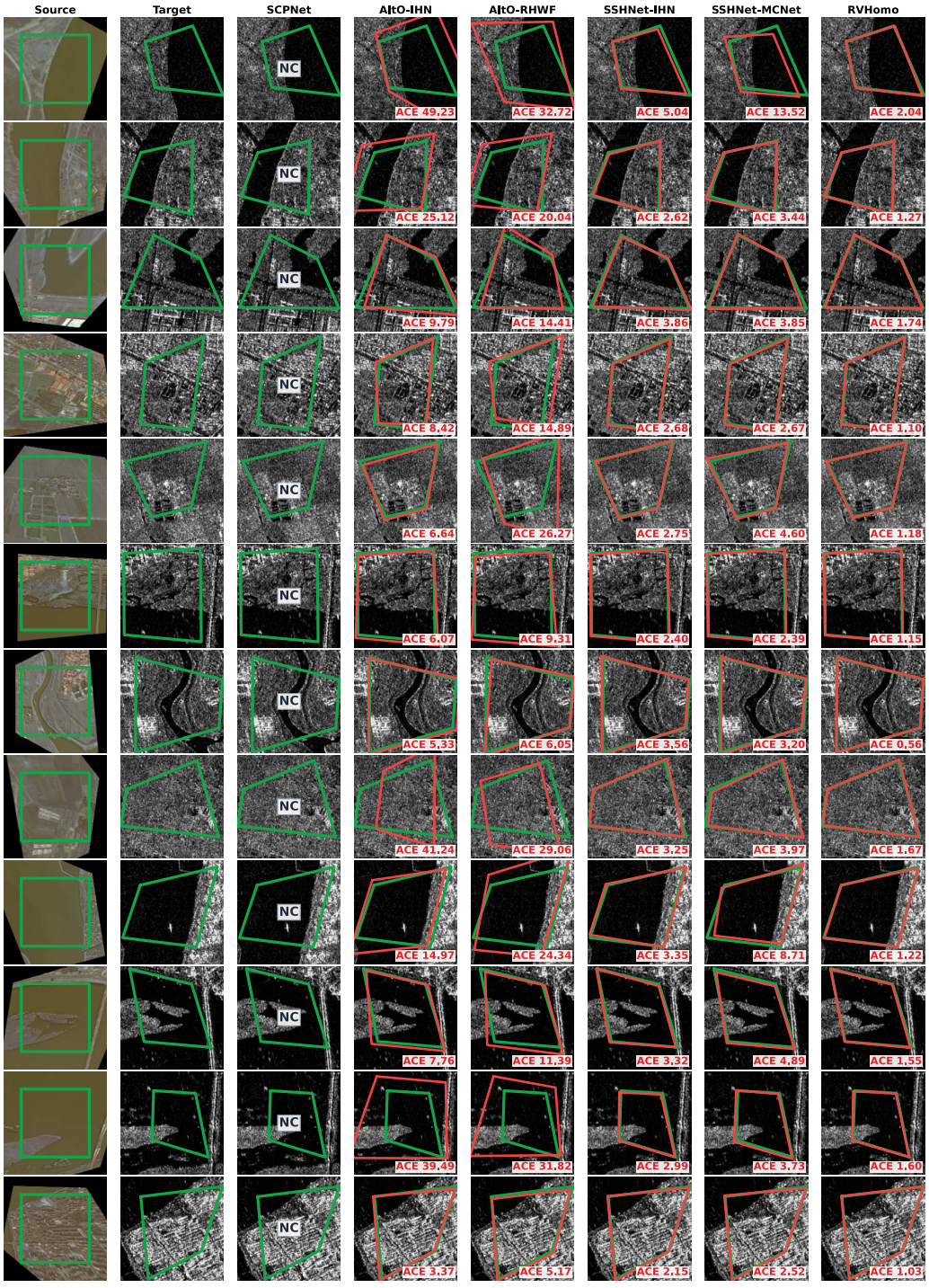


Figure E.2: Qualitative comparison of different methods on OPT-SAR.



Figure E.3: Qualitative comparison of different methods on FSD.

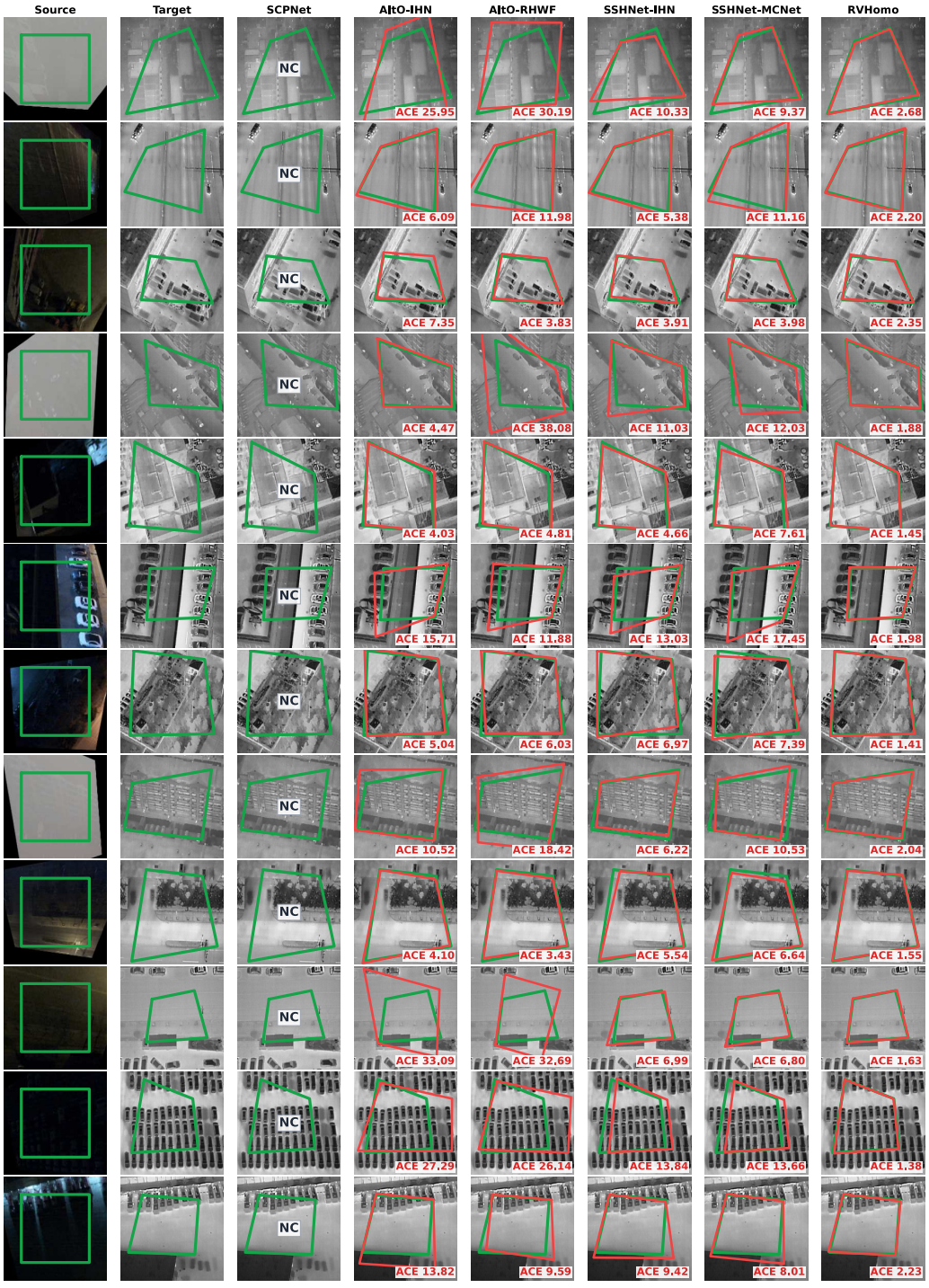


Figure E.4: Qualitative comparison of different methods on VIS-IR.