TAILOR-MADE FOR ONE: LEARNING FROM EACH TEST PAIR FOR ZERO-SHOT CROSS-MODAL HOMOGRAPHY ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-modal homography methods typically rely on training data specific to the cross-modal combination encountered at deployment. Existing zero-shot methods alleviate this dependency by learning modality-invariant structural representations, but still treat each test pair as a passive input, leaving its pair-specific information unexplored. We introduce Tailor-Made for One (TMO), a zero-shot framework that turns each unlabeled test pair into an adaptation signal. The central challenge is that direct reconstruction-based adaptation can conflate appearance differences with geometric misalignment, allowing modality translation to distort the scene structure. TMO addresses this ambiguity by synthesizing prediction-centered training pairs with varying source geometries around the current alignment and requiring adaptation to remain effective across these pairs. This provides pair-specific supervision that refines appearance transfer while preserving its compatibility with a frozen geometric prior. On the three benchmarks shared with prior zero-shot work, TMO achieves state-of-the-art performance before adaptation. Across six benchmarks, adaptation tailored to each test pair further reduces MACE by 16.9%–50.8%. These results show that an unlabeled test pair is not merely an input to a transferable model, but a source of supervision for zero-shot geometric estimation.

1 INTRODUCTION

Cross-modal homography estimation recovers a planar projective transformation that aligns overlapping images from different modalities (Zhang et al., 2025). Mapping heterogeneous observations to a shared frame provides the geometric foundation for cross-sensor image fusion (Wu et al., 2025; Ying et al., 2021), localization (Zhong et al., 2026; Wang et al., 2023), multimodal scene understanding (Chen et al., 2025; Wang et al., 2025), and other geometry-dependent vision tasks (Cai & Yang, 2025; Ding et al., 2025).

Most learning-based approaches address cross-modal appearance ambiguity through modality-specific training. Supervised methods rely on paired images with geometric annotations (Cao et al., 2023; Zhu et al., 2024; Cao et al., 2022; DeTone et al., 2016), whereas unsupervised methods replace such annotations with reconstruction or consistency objectives (Yu et al., 2025; Song et al., 2024; Zhang et al., 2024; Liu et al., 2022b). Despite differing supervision, both paradigms typically require real training pairs from the target modality combination. Zero-shot methods instead learn transferable structural representations from generic images, avoiding target-modality pairs during offline training (You et al., 2026). Synthetic multimodal training further broadens this appearance coverage (Ren et al., 2025), but may not capture the sensor-specific statistics and scene-dependent discrepancies of every test pair. With fixed inference parameters, the observed pair informs prediction but cannot specialize the model through learning. Test-time adaptation addresses this distinct limitation by learning from deployment data, as explored in multimodal action recognition (Xiong et al., 2024), depth completion (Park et al., 2024), and vision-language models (Huang et al., 2026).

Yet extending test-time adaptation (Gou et al., 2024; Zhou et al., 2024; Wang et al., 2021; Sun et al., 2020) to cross-modal homography estimation remains largely unexplored. A single unlabeled test pair provides no direct geometric supervision, while modality differences and spatial misalignment

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.

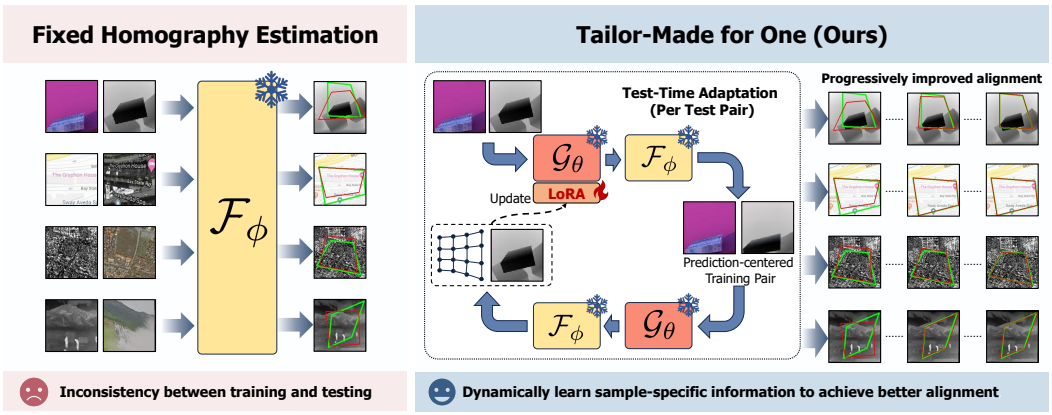

Figure 1: Conventional inference (left) and TMO (right) for zero-shot homography estimation. Differences between training and test data can impair inference accuracy. TMO adapts a modality translator ($\mathcal{G}_\theta$) to each test pair to learn cross-modal appearance relationships while a homography estimator ($\mathcal{F}_\phi$) provides geometric guidance, enabling progressively improved alignment.

make reliable correspondences difficult to establish. Without ground truth to verify the inferred alignment, adaptation may reinforce erroneous estimates over successive updates and drift away from the correct geometry. The challenge is therefore to derive reliable supervision from the test pair while distinguishing appearance differences from geometric misalignment.

This leads to our central question: *How can a single unlabeled test pair refine a strong zero-shot alignment without conflating appearance adaptation with geometric correction?* Our insight is to learn appearance transfer across varying local geometries under a fixed geometric estimator. Prediction-centered data synthesis produces prediction-centered training pairs that share the same appearance relationship but differ in source geometry. Applying supervision to these pairs discourages dependence on one fixed spatial configuration and encourages appearance transfer that remains useful across perturbations. This regularizes adaptation against absorbing residual misalignment, while re-estimation refreshes the geometric anchor as the translation changes.

We realize this idea through Tailor-Made for One (TMO). Figure 1 highlights the contrast: conventional inference with fixed parameters cannot learn from unseen modality differences, whereas TMO adapts to them using each test pair. TMO separates modality translation from homography estimation. Both networks are pretrained on generic RGB images with synthetic appearance variations and known projective perturbations. At test time, the translator learns the appearance relationship between the two modalities through reconstruction and feature supervision. The homography estimator remains frozen to preserve its pretrained ability to establish robust geometric correspondences and guide these updates. To make adaptation efficient while retaining pretrained knowledge, we update only low-rank adaptation (LoRA) (Hu et al., 2022) modules in the translator, keeping its backbone frozen. The adapters are reset between pairs, allowing the same pretrained model to accommodate different modality combinations.

Our main contributions are summarized as follows:

- We introduce TMO, a framework for zero-shot cross-modal homography estimation that learns from each unlabeled test pair. It adapts appearance transfer under a frozen geometric prior, allowing a shared pretrained model to specialize to the observed modality differences without geometric labels.
- We develop prediction-centered data synthesis with iterative anchor refresh to generate training pairs with varying source geometries. Learning a shared appearance relationship across these pairs discourages the translator from absorbing residual misalignment, addressing the ambiguity between appearance adaptation and geometric correction.
- With pretraining only on generic RGB images, TMO outperforms the reported zero-shot baseline on three shared benchmarks before adaptation. Learning from each test pair further reduces MACE by 16.9%–50.8% across six benchmarks, with gains also observed using alternative translators and estimators.

2 Related Work

Supervised Homography Estimation. Supervised methods learn to estimate homographies from image pairs with known transformations. DHN (DeTone et al., 2016) introduced direct CNN regression of four-point displacements from synthetically warped pairs. MHN (Le et al., 2020) uses multiscale estimation. DLKFM couples learned modality-invariant features with Lucas–Kanade optimization (Zhao et al., 2021), while LocalTrans uses local attention within a multiscale architecture for cross-resolution matching (Shao et al., 2021). Recent methods iteratively refine homography estimates using learned update operators (Cao et al., 2022), homography-guided image warping and attention (Cao et al., 2023), or multiscale correlation searching (Zhu et al., 2024). GFNet estimates homographies from flow predicted on a coarse grid (Zhang et al., 2025). Cross-modal evaluations commonly involve training or fine-tuning on pairs from the evaluated modality combination with geometric supervision. Their standard inference procedures keep network parameters fixed rather than learning from each unlabeled test pair.

Unsupervised Homography Estimation. Unsupervised methods require no geometric annotations for the observed pairs. UDHN uses differentiable warping and photometric consistency (Nguyen et al., 2018). CA-UDHN combines learned masks with a feature-based triplet loss (Zhang et al., 2020), while biHomE uses bidirectional perceptual supervision with frozen features (Koguciuk et al., 2021). BasesHomo learns weights for predefined homography flow bases and applies low-rank feature projection (Liu et al., 2022b). For cross-modal estimation, SCPNet combines intra-modal self-supervision with consistent feature projection (Zhang et al., 2024); AltO alternates geometric alignment and modality-invariant feature learning (Song et al., 2024); and SSHNet separates modality transfer and homography estimation through split optimization (Yu et al., 2025). Despite requiring no geometric annotations, these methods typically rely on substantial collections of image pairs from the target modality combination for training.

Zero-Shot Generalization. Zero-shot cross-modal homography estimation seeks to generalize without training pairs from the target modality combination. CCNet is trained on synthetic pairs with diverse appearances and preserved structure (You et al., 2026). Related correspondence work explores complementary routes: MINIMA constructs synthetic multimodal matching data (Ren et al., 2025), ZeroCo extracts correspondences from cross-attention maps in pretrained cross-view completion models (An et al., 2025), and GIM learns generalizable image matchers through self-training on diverse internet videos (Shen et al., 2024). Despite improved generalization, pretraining data may not capture the sensor characteristics and appearance statistics of every test pair. With parameters fixed at inference, the models cannot learn from the observed pair to narrow this distribution gap.

3 Method

3.1 Overall Framework

As illustrated in Figure 2, TMO combines a conditional modality translator $\mathcal{G}_\theta$ with a homography estimator $\mathcal{F}_\phi$ for zero-shot cross-modal homography estimation. The translator adapts appearance, while the estimator provides geometric guidance. Both are pretrained solely on generic RGB images with synthetic appearance variations and known projective perturbations. Following CCNet (You et al., 2026), this zero-shot setting excludes real target-modality pairs from offline training.

At test time, both backbones are frozen and only translator low-rank adaptation (LoRA) (Hu et al., 2022) parameters ψ are updated independently for each pair. Given $\mathcal{Q} = (\mathbf{X}_s, \mathbf{X}_t)$, with $\mathbf{X}_s, \mathbf{X}_t \in \mathbb{R}^{H\times W\times 3}$, the forward pass is

$$\mathbf{Z} = \mathcal{G}_{\theta,\psi}(\mathbf{X}_s, \mathbf{X}_t), \qquad \widehat{\Delta\mathbf{P}} = \mathcal{F}_\phi(\mathbf{Z}, \mathbf{X}_t). \tag{1}$$

The translator is trained to make $\mathbf{Z}$ match the target appearance while retaining the source coordinate frame. The corner displacements $\widehat{\Delta\mathbf{P}} \in \mathbb{R}^{4\times 2}$ are converted by the Direct Linear Transform (DLT) into $\widehat{\mathbf{H}} \in \mathbb{R}^{3\times 3}$, which maps the source coordinate frame to the target coordinate frame.

This prediction anchors adaptation, but reconstruction under a single inaccurate alignment can encourage the translator to absorb residual geometric errors. TMO therefore generates prediction-centered training pairs by sampling local perturbations around the anchor while keeping the target

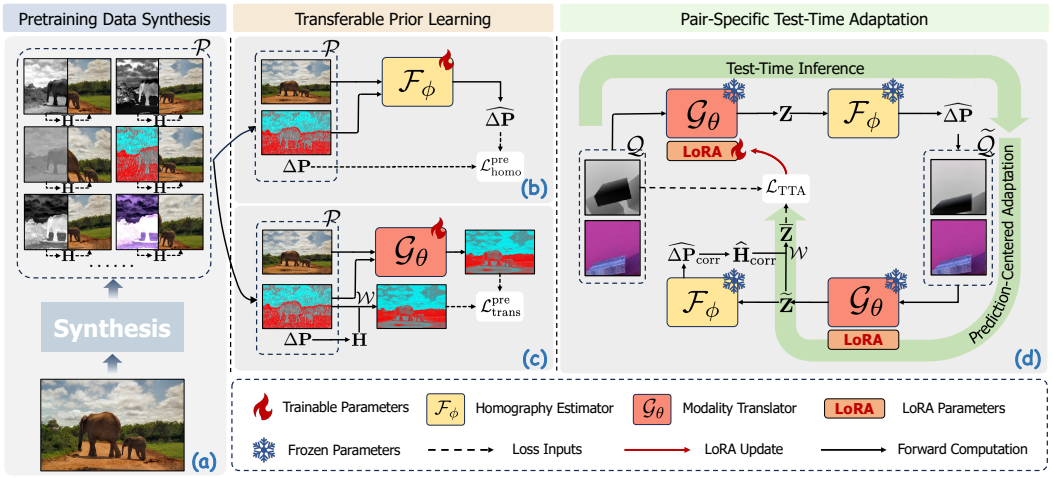

Figure 2: Overview of TMO. (a) Generic RGB images yield synthetic cross-appearance pairs with known geometry. (b) Corner displacements supervise homography pretraining. (c) Known warps provide aligned targets for translator pretraining. (d) At test time, the current prediction anchors synthesis of a perturbed pair. The translator transfers appearance, and the frozen estimator predicts a correction warp for reconstruction and feature supervision. Only translator LoRA parameters are updated; both backbones remain frozen. Re-estimation on the original pair refreshes the anchor.

fixed. All training pairs derive from the same observed pair, so adaptation addresses the same cross-modal appearance gap across different source geometries. Optimizing a shared translator across these pairs discourages alignment-specific geometric corrections. At each iteration, the frozen estimator predicts a detached correction homography for each training pair. This homography warps the translated source into the target frame, where reconstruction and homography-aware feature losses compare it with the target using pixels and frozen estimator features, respectively. After updating ψ, re-estimation on the original pair refreshes the anchor. The final prediction uses a translator *tailor-made for one* test pair; its pair-specific LoRA parameters are discarded after inference.

3.2 Learning a Transferable Prior

Pretraining learns a transferable prior from synthetic pairs with shared, known geometry. The estimator learns appearance-robust geometry for zero-shot prediction and adaptation guidance, while the translator learns geometry-preserving appearance transfer.

Pretraining Data Synthesis. We construct synthetic training pairs from natural RGB images by independently simulating appearance and geometric variations. Given a randomly cropped image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$, we generate $\mathbf{y} = \mathcal{S}_m(\mathbf{x}; \boldsymbol{\xi})$, where $\mathcal{S}_m$ is sampled online from a family of appearance transformations and $\boldsymbol{\xi}$ denotes its random parameters. The family includes channel and intensity remapping, nonlinear contrast adjustment, texture enhancement, smoothing, resolution degradation, quantization, and stochastic noise. These transformations apply no additional global projective warp, so $\mathbf{x}$ and $\mathbf{y}$ retain the same coordinate frame.

To introduce a known geometric displacement, we independently sample each corner displacement as $\Delta\mathbf{p}_i \sim \mathrm{Unif}(\{-\rho, \dots, \rho\}^2)$ for $i \in \{1, \dots, 4\}$ and stack them as $\Delta\mathbf{P} = [\Delta\mathbf{p}_1, \dots, \Delta\mathbf{p}_4]^\top \in \mathbb{R}^{4\times 2}$, where $\rho = 32$ pixels. Let $\mathbf{p}_i \in \mathbb{R}^2$ denote the reference corners ordered as top-left, top-right, bottom-left, and bottom-right. The corner matrix and corresponding homography are

$$\mathbf{P} = [\mathbf{p}_1, \dots, \mathbf{p}_4]^\top \in \mathbb{R}^{4\times 2}, \qquad \mathbf{H} = \mathrm{DLT}(\mathbf{P}, \mathbf{P} + \Delta\mathbf{P}). \tag{2}$$

The same conversion applies to predicted corner displacements. Under the adopted inverse-warping convention, we synthesize

$$\mathcal{P} = (\mathbf{I}_s, \mathbf{I}_t) = \big(\mathcal{W}(\mathbf{x}; \mathbf{H}^{-1}), \mathbf{y}\big), \tag{3}$$

where $\mathcal{W}(\cdot; \mathbf{H}^{-1})$ generates the geometrically perturbed source image, such that $\mathbf{H}$ aligns it with the target frame. The corresponding ground-truth four-point displacement is $\Delta\mathbf{P}$. The same synthetic pair $\mathcal{P}$ is used in both pretraining stages below: $\Delta\mathbf{P}$ supervises the homography estimator,

while the known warp establishes aligned translation targets. This provides exact geometric control under diverse appearance variations without requiring real paired cross-modal data. Additional implementation details of the synthesis procedure are provided in Appendix A.3.

Pretraining for Homography Estimation. We pretrain the homography estimator $\mathcal{F}_\phi$ on $\mathcal{P}$. Because $\mathcal{S}_m$ is sampled online, these pairs expose the estimator to diverse appearance discrepancies while retaining the known displacement $\Delta\mathbf{P}$. By requiring consistent homography prediction across such variations, this stage encourages $\mathcal{F}_\phi$ to learn representations that are invariant to appearance yet sensitive to projective geometry. Denoting the estimated displacement by $\widehat{\Delta\mathbf{P}} = \mathcal{F}_\phi(\mathbf{I}_s, \mathbf{I}_t)$, we optimize the estimator using direct supervision from the known displacement:

$$\mathcal{L}_{\text{homo}}^{\text{pre}} = \left\| \widehat{\Delta\mathbf{P}} - \Delta\mathbf{P} \right\|_1 . \tag{4}$$

After pretraining, we freeze $\mathcal{F}_\phi$ and use its appearance-invariant, geometry-sensitive feature encoder as a fixed geometric teacher for modality translation.

Pretraining for Modality Translation. We pretrain a conditional translator $\mathcal{G}_\theta(\mathbf{I}, \mathbf{C})$ to transfer the appearance specified by the condition image $\mathbf{C}$ to the source image $\mathbf{I}$ while preserving the source geometry. The source is the explicit carrier of spatial information, whereas the condition supplies a spatially unordered appearance representation (Appendix A.2). We treat this separation, together with aligned reconstruction and feature supervision, as establishing a basic capacity for geometry-preserving appearance transfer rather than guaranteeing it. During adaptation (Section 3.3), prediction-centered perturbations further reinforce this behavior by discouraging the translator from absorbing residual geometric misalignment.

Using the synthetic pair $\mathcal{P}$ defined above, we train the translator in both directions: $\widehat{\mathbf{I}}_s = \mathcal{G}_\theta(\mathbf{I}_s, \mathbf{I}_t)$ preserves the geometry of $\mathbf{I}_s$ while adopting the appearance of $\mathbf{I}_t$, whereas $\widehat{\mathbf{I}}_t = \mathcal{G}_\theta(\mathbf{I}_t, \mathbf{I}_s)$ preserves the geometry of $\mathbf{I}_t$ while adopting the appearance of $\mathbf{I}_s$. Their targets are $\mathbf{I}_t^H = \mathcal{W}(\mathbf{y}; \mathbf{H}^{-1})$ and $\mathbf{x}$, respectively. The reconstruction loss is

$$\mathcal{L}_{\text{reg}}^{\text{pre}} = \frac{1}{2}\left(\left\| \widehat{\mathbf{I}}_s - \mathbf{I}_t^H \right\|_1 + \left\| \widehat{\mathbf{I}}_t - \mathbf{x} \right\|_1 \right). \tag{5}$$

Pixel regression alone does not ensure that the translated outputs retain the representations required for geometric matching. We therefore use the frozen homography encoder as a feature teacher. Let $\mathbf{T}_s = \mathbf{I}_t^H$ and $\mathbf{T}_t = \mathbf{x}$ denote the aligned targets, and let f_ϕ^ℓ be its feature at scale ℓ. We minimize the cosine distance between each translated output and its aligned target, with sg denoting stop-gradient:

$$\mathcal{L}_{\text{feat}}^{\text{pre}} = \frac{1}{2L} \sum_{d\in\{s,t\}} \sum_{\ell=1}^{L} \left[1 - \cos\Big(f_\phi^\ell(\widehat{\mathbf{I}}_d), \mathrm{sg}\big[f_\phi^\ell(\mathbf{T}_d) \big] \Big) \right]. \tag{6}$$

The encoder parameters remain fixed, but gradients through $f_\phi^\ell(\widehat{\mathbf{I}}_d)$ update the translator, encouraging its outputs to preserve homography-relevant structure. The translator therefore minimizes the weighted objective $\mathcal{L}_{\text{trans}}^{\text{pre}} = \lambda_{\text{reg}}^{\text{pre}}\mathcal{L}_{\text{reg}}^{\text{pre}} + \lambda_{\text{feat}}^{\text{pre}}\mathcal{L}_{\text{feat}}^{\text{pre}}$.

3.3 Pair-Specific Iterative Adaptation

At test time, we treat each unseen cross-modal pair $\mathcal{Q} = (\mathbf{X}_s, \mathbf{X}_t)$ as an independent adaptation task. LoRA augments each selected translator layer $k \in \mathcal{K}$ as

$$\mathcal{O}_k(\mathbf{h}_k) = \mathcal{O}_{0,k}(\mathbf{h}_k) + \frac{\alpha}{r}\mathcal{B}_k\big(\mathcal{A}_k(\mathbf{h}_k)\big), \tag{7}$$

where $\mathbf{h}_k$ is the input activation, $\mathcal{O}_{0,k}$ is the frozen pretrained operator, and $\mathcal{A}_k$ and $\mathcal{B}_k$ project to and from an r-dimensional space. The factor α/r scales the update, and ψ collects the projection parameters across layers. For each new pair, fresh adapters $\psi^{(0)}$ produce a zero update, so $\mathcal{G}_{\theta,\psi^{(0)}} = \mathcal{G}_\theta$; both pretrained backbones remain fixed, and $\mathcal{F}_\phi$ stays in evaluation mode. At iteration n, we first obtain

$$\mathbf{Z}^{(n)} = \mathcal{G}_{\theta,\psi^{(n)}}(\mathbf{X}_s, \mathbf{X}_t), \qquad \widehat{\Delta\mathbf{P}}^{(n)} = \mathrm{sg}\Big[\mathcal{F}_\phi(\mathbf{Z}^{(n)}, \mathbf{X}_t) \Big], \tag{8}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. Although the estimator is frozen, its prediction evolves with the adapted translation and provides a new geometric anchor at each iteration.

Prediction-centered data synthesis. A single test pair exposes only one coupling between cross-modal appearance and geometry. After warping the source by the current prediction, any residual misalignment remains entangled with the modality difference. Reconstruction from this single configuration may therefore encourage the translator to encode residual geometric error as part of the appearance transfer. To vary geometry without changing the observed cross-modal appearance relationship, we synthesize a local neighborhood around the current prediction while keeping the target fixed. Specifically, the current homography $\widehat{\mathbf{H}}^{(n)}$ is induced by $\widehat{\Delta\mathbf{P}}^{(n)}$. We then sample a random four-point residual $\delta\mathbf{P}_r^{(n)}$, with $\delta\mathbf{p}_{r,i}^{(n)} \sim \mathrm{Unif}(\{-\eta, \dots, \eta\}^2)$, and construct its homography $\mathbf{R}^{(n)}$. The perturbed source is

$$\widetilde{\mathbf{X}}_s^{(n)} = \mathcal{W}\Big(\mathbf{X}_s; (\mathbf{R}^{(n)})^{-1}\widehat{\mathbf{H}}^{(n)}\Big). \tag{9}$$

Under our mapping convention, the composition $(\mathbf{R}^{(n)})^{-1}\widehat{\mathbf{H}}^{(n)}$ first applies the current source-to-target alignment and then the inverse residual perturbation. Thus, if the current estimate is exact, $\mathbf{R}^{(n)}$ alone aligns the perturbed source with the target. We call $\widetilde{\mathcal{Q}}^{(n)} = (\widetilde{\mathbf{X}}_s^{(n)}, \mathbf{X}_t)$ a *prediction-centered training pair*. It retains the observed modalities and their appearance relationship while presenting a different source geometry. Resampling the residual yields multiple such configurations from the same test pair. Optimizing a shared translator across them makes a fixed target-aligned output or configuration-specific spatial correction less effective, providing pair-specific supervision that encourages appearance transfer to retain the geometry of each perturbed source.

Target-image reconstruction. Given $\widetilde{\mathcal{Q}}^{(n)}$, the translator first maps the perturbed source toward the target appearance. The frozen estimator then predicts a detached correction and uses it to align the translated source:

$$\widetilde{\mathbf{Z}}^{(n)} = \mathcal{G}_{\theta,\psi^{(n)}}\big(\widetilde{\mathbf{X}}_s^{(n)}, \mathbf{X}_t\big), \quad \widehat{\Delta\mathbf{P}}_{\mathrm{corr}}^{(n)} = \mathrm{sg}\Big[\mathcal{F}_\phi\big(\widetilde{\mathbf{Z}}^{(n)}, \mathbf{X}_t\big)\Big], \quad \overline{\mathbf{Z}}^{(n)} = \mathcal{W}\big(\widetilde{\mathbf{Z}}^{(n)}; \widehat{\mathbf{H}}_{\mathrm{corr}}^{(n)}\big), \tag{10}$$

where $\widehat{\mathbf{H}}_{\mathrm{corr}}^{(n)}$ is induced by $\widehat{\Delta\mathbf{P}}_{\mathrm{corr}}^{(n)}$ (Eq. 2). The stop-gradient treats the predicted correction as fixed during each translator update. The resulting homography maps the translated perturbed source to the target frame. Let $\mathbf{M}^{(n)}$ denote the validity mask propagated through both the prediction-centered source warp and the correction warp. We reconstruct the observed target over this valid support:

$$\mathcal{L}_{\mathrm{rec}}^{(n)} = \Big\|\mathbf{M}^{(n)} \odot \big(\overline{\mathbf{Z}}^{(n)} - \mathbf{X}_t\big)\Big\|_1. \tag{11}$$

This objective learns the pair-specific appearance relationship without requiring the ground-truth homography.

Homography-aware feature alignment. Pixel-wise reconstruction alone may produce a visually plausible translation that is incompatible with the pretrained geometric prior. To preserve the representations used for homography estimation, we further align the multi-scale features $\{f_\phi^\ell\}_{\ell=1}^L$ extracted by the frozen estimator:

$$\mathcal{L}_{\mathrm{feat}}^{(n)} = \frac{1}{L}\sum_{\ell=1}^{L}\Big[1 - \cos\Big(f_\phi^\ell(\mathbf{M}^{(n)} \odot \overline{\mathbf{Z}}^{(n)}), f_\phi^\ell(\mathbf{M}^{(n)} \odot \mathbf{X}_t)\Big)\Big]. \tag{12}$$

We broadcast $\mathbf{M}^{(n)}$ over image channels and mask both inputs before feature extraction, suppressing invalid content introduced by warping. At each scale, we average the cosine distance over the feature grid and then across the L scales. The frozen encoder thereby provides a stable, homography-aware feature space for supervision. Although its parameters remain fixed, gradients through the translated input are retained, allowing this loss to update the translator toward representations compatible with the geometric estimator.

Combining appearance reconstruction with homography-aware feature alignment, we minimize $\mathcal{L}_{\mathrm{TTA}}^{(n)} = \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}}^{(n)} + \lambda_{\mathrm{feat}}\mathcal{L}_{\mathrm{feat}}^{(n)}$, where λ_{rec} and λ_{feat} balance the two terms. Both losses update only $\psi^{(n)}$, yielding $\psi^{(n+1)}$ for the next iteration while preserving both the pretrained translator backbone and the geometric prior. After N iterations, we repeat the forward pass with the final adapted LoRA parameters to obtain $\widehat{\Delta\mathbf{P}}^*$. No ground-truth displacement of $\mathcal{Q}$ is used for optimization, and the instance-specific LoRA parameters are discarded before processing the next test pair. This closed loop tailors the translated representation to each test pair without modifying either pretrained network backbone.

Table 1: Cross-modal homography comparison in MACE (pixels; lower is better). NC denotes non-convergence and −− unavailable results. TMO_0 denotes the pretrained model, while TMO_{300} performs 300 unlabeled adaptation steps for each image pair. Result sources and evaluation protocols are detailed in Appendix A.1. Within each supervision type, the best and second-best results are highlighted using dark and light gray-blue backgrounds, respectively.

Type	Method	OPT-SAR	GoogleMap	GoogleEarth	DPDN	RGB-NIR	M3FD
Unsupervised	UDHN	NC	22.84	21.20	NC	23.43	NC
	CA-UDHN	24.76	24.61	21.32	24.99	24.12	24.88
	BasesHomo	26.72	24.49	25.12	27.33	24.41	24.85
	SCPNet	NC	4.35	2.67	NC	4.78	NC
	AltO	8.93	4.73	1.91	18.87	3.11	NC
	SSHNet	2.94	1.23	1.51	1.12	1.66	3.04
Supervised	DHN	8.27	5.20	6.42	4.92	11.88	5.21
	MHN	5.59	2.69	5.24	2.95	5.52	5.88
	LocalTrans	20.68	3.22	1.52	1.79	1.72	13.39
	IHN	1.67	0.92	0.80	1.17	1.63	1.25
	RHWF	1.48	0.71	0.65	1.06	1.07	1.25
	MCNet	1.24	1.02	0.93	0.85	0.62	1.07
Zero-shot	CCNet	–	4.38	1.40	–	4.46	–
	TMO_0	13.66	3.29	1.30	8.63	1.36	5.85
	TMO_{300}	8.00	1.62	1.08	6.10	0.69	4.09

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate on the official test or validation splits of OPT-SAR (Li et al., 2022), GoogleMap and GoogleEarth (Zhao et al., 2021), DPDN (Riegler et al., 2016), RGB-NIR (Brown & Süsstrunk, 2011), and M3FD (Liu et al., 2022a). Offline pretraining uses only MS COCO (Lin et al., 2014). Modality pairs, split sizes, and evaluation protocols appear in Appendix A.1.

Implementation. Unless otherwise specified, TMO uses MCNet (Zhu et al., 2024) as its homography estimator. The conditional translator architecture is detailed in Appendix A.2, and pretraining and test-time adaptation settings are provided in Appendix A.3.

Baselines. We compare with six unsupervised methods: UDHN (Nguyen et al., 2018), CA-UDHN (Zhang et al., 2020), BasesHomo (Liu et al., 2022b), SCPNet (Zhang et al., 2024), AltO (Song et al., 2024), and SSHNet (Yu et al., 2025); six supervised methods: DHN (DeTone et al., 2016), MHN (Le et al., 2020), LocalTrans (Shao et al., 2021), IHN (Cao et al., 2022), RHWF (Cao et al., 2023), and MCNet (Zhu et al., 2024); and the zero-shot method CCNet (You et al., 2026). Evaluation protocols and result sources are detailed in Appendix A.1.

Metrics. We report Mean Average Corner Error (MACE) (DeTone et al., 2016), the Euclidean corner error averaged over four corners and all test pairs (pixels; lower is better). Ground-truth homographies are used only for evaluation, never for adaptation.

4.2 Comparisons with Existing Approaches

Standalone zero-shot performance. Table 1 first establishes the quality of the pretrained model. TMO_0 achieves MACE of 3.29, 1.30, and 1.36 on GoogleMap, GoogleEarth, and RGB-NIR, improving over the existing zero-shot baseline by 24.9%, 7.1%, and 69.5%, respectively. It therefore sets a new zero-shot state of the art on these three benchmarks without any test-time update. The learned composition is effective as a standalone estimator.

Adaptation beyond a strong initialization. Starting from the same pretrained weights, TMO_{300} reduces these errors to 1.62, 1.08, and 0.69, corresponding to further reductions of 50.8%, 16.9%, and 49.3%. GoogleEarth and RGB-NIR are particularly informative: their initial errors are already near one pixel, yet adaptation improves both. Across all six benchmarks, MACE decreases by 16.9%–50.8%. The TMO_0–TMO_{300} comparison thus separates the contribution of instance-specific learning from the pretrained model's existing advantage.

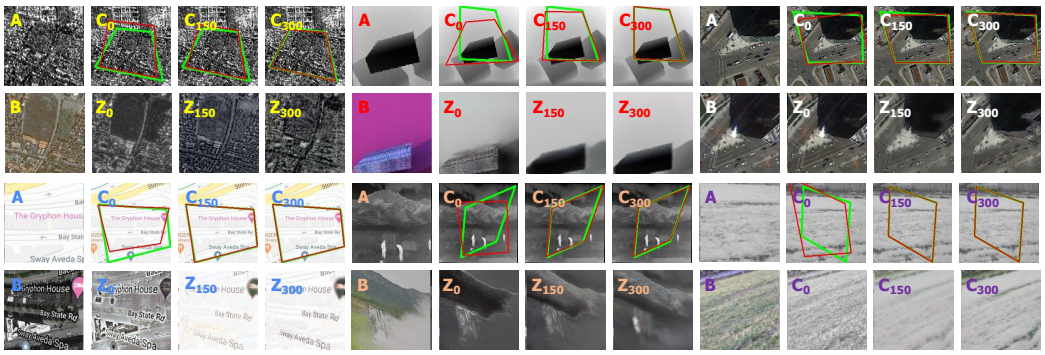

Figure 3: Representative adaptation results on six benchmarks, one per block. **A** and **B** are the unaligned input images; **C** shows alignment comparisons, and **Z** shows modality-translated images. Subscripts denote adaptation steps (0, 150, and 300). Closer overlap of the red and green outlines indicates better alignment.

Table 2: Evolution of MACE during instance-specific test-time adaptation. Iteration 0 denotes the initial zero-shot prediction, and iteration 300 denotes the final adapted prediction. Results are reported as MACE in pixels; lower is better.

Iteration	OPT-SAR	GoogleMap	GoogleEarth	DPDN	RGB-NIR	M3FD
0	13.66	3.29	1.30	8.63	1.36	5.85
50	10.67	2.67	1.25	7.15	0.95	5.06
100	9.42	2.33	1.17	6.69	0.87	4.57
150	8.68	1.93	1.12	6.36	0.87	4.25
200	8.24	1.75	1.10	6.26	0.80	4.20
250	8.01	1.63	1.07	6.13	0.73	4.13
300	8.00	1.62	1.08	6.10	0.69	4.09

Comparison across supervision regimes. TMO_{300} outperforms the best listed unsupervised results on GoogleEarth and RGB-NIR by 28.5% and 58.4%, respectively. On GoogleMap, M3FD, OPT-SAR, and DPDN, it is surpassed only by SSHNet among the listed unsupervised baselines. Notably, on OPT-SAR and DPDN, which present pronounced modality gaps, TMO_{300} surpasses AltO, the second-best listed unsupervised baseline, reducing MACE by 10.4% and 67.7%, respectively. It also outperforms at least one supervised baseline on five benchmarks and approaches the best supervised RGB-NIR result (0.69 vs. 0.62). Unlike target-trained baselines, TMO uses no offline target-modality pairs and obtains target-specific supervision solely from the current unlabeled pair through test-time optimization.

Progressive appearance and alignment refinement. Figure 3 visualizes the coupled progress of modality translation and geometric alignment. As adaptation proceeds, the translated output **Z** generated from modality **B** approaches the appearance of modality **A**. Meanwhile, the red and green outlines in **C** overlap increasingly closely, indicating improved alignment. The quantitative trajectories in Table 2 corroborate this trend: MACE decreases rapidly at early iterations and then gradually saturates on all benchmarks. From iteration 0 to 300, the mean of the six dataset-wise relative reductions is 36.3%, and by iteration 200 every benchmark has realized at least 83.6% of its total 300-step gain. Between iterations 250 and 300, dataset-level MACE changes by at most 0.04 pixels; five benchmarks continue to improve, while GoogleEarth exhibits only a 0.01-pixel fluctuation after reaching 1.07. These visual and quantitative trends show that TMO refines appearance transfer and geometric estimation, with aggregate errors approaching stable plateaus under the common 300-step schedule. Further benefit–step and per-pair distribution analyses are provided in Appendices B.1 and B.2.

4.3 Ablation

Architectural components. Table 3 varies the estimator with our translator fixed, and the translator with MCNet fixed. With RHWF, IHN, and MCNet, adaptation reduces MACE on every benchmark by 18.0%–38.8%, 15.8%–43.1%, and 16.9%–50.8%, respectively. U-Net also improves on all six

Table 3: Component ablations in TMO. Entries report MACE at iterations 0/300 (pixels; lower is better), with the other component fixed. *Default* denotes MCNet and our translator in the estimator and translator blocks, respectively. U-Net uses the architecture from SSHNet (Yu et al., 2025).

Component	Variant	OPT-SAR	GoogleMap	GoogleEarth	DPDN	RGB-NIR	M3FD
Homography estimator	RHWF	11.16/7.35	3.38/2.11	1.61/1.32	7.82/6.15	1.60/0.98	5.93/4.58
	IHN	12.62/7.73	4.62/2.63	1.81/1.51	9.11/7.67	2.03/1.24	7.70/5.78
	Default	13.66/8.00	3.29/1.62	1.30/1.08	8.63/6.10	1.36/0.69	5.85/4.09
Modality translator	U-Net	13.97/10.34	5.14/2.74	1.85/1.78	8.22/6.63	2.21/1.51	6.58/6.49
	Default	13.66/8.00	3.29/1.62	1.30/1.08	8.63/6.10	1.36/0.69	5.85/4.09

Table 4: Ablations of adaptation objectives (upper; prediction-centered perturbations) and perturbation strategies (lower; both losses enabled). No adaptation is the shared zero-shot reference. MACE is in pixels; lower is better.

Setting	$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{feat}}$	OPT-SAR	GoogleMap	GoogleEarth	DPDN	RGB-NIR	M3FD
No adaptation	Off	Off	13.66	3.29	1.30	8.63	1.36	5.85
Reconstruction only	On	Off	8.71	2.32	1.10	6.63	0.85	5.30
Feature only	Off	On	8.08	1.65	1.09	7.20	0.68	4.23
Full objective	On	On	8.00	1.62	1.08	6.10	0.69	4.09
No residual perturbation	On	On	9.52	1.96	1.21	8.21	1.01	5.34
Direct (origin-centered)	On	On	8.39	1.88	1.07	6.65	0.96	4.29
Prediction-centered (default)	On	On	8.00	1.62	1.08	6.10	0.69	4.09

benchmarks, by 26.0%, 46.7%, 3.8%, 19.3%, 31.7%, and 1.4%, although its gains on GoogleEarth and M3FD are modest. Our default translator achieves lower final errors on every benchmark and lower initial errors on five of six; U-Net has a slightly lower DPDN initialization (8.22 vs. 8.63). These results distinguish the influence of component design on initial accuracy from the adaptability shared by the tested configurations. The two Default rows represent the same complete TMO model.

Adaptation objectives. Table 4 compares target-image reconstruction and homography-aware feature alignment. Both improve over the shared initialization, reducing MACE by 9.4%–37.5% and 16.2%–50.0%, respectively, which confirms that each provides an effective adaptation signal. Their combination achieves the lowest error on five of the six benchmarks, supporting the complementary roles of pixel-level appearance reconstruction and homography-aware feature supervision. The only exception is RGB-NIR, where feature alignment alone is lower by just 0.01 pixels (0.68 vs. 0.69), a difference comparable to the small late-stage error fluctuations. Overall, jointly applying the two losses yields the strongest and most consistent performance across datasets.

Perturbation strategies. The lower block of Table 4 fixes both losses and varies the synthesis geometry. Without residual perturbations, the current prediction supplies the synthesis warp; no additional geometric neighborhood is sampled. Adaptation may then overfit the pair and absorb residual misalignment into the translation. Although this setting improves over the shared initialization on all benchmarks, it remains worse than both residual-perturbation strategies on every benchmark. Direct, origin-centered perturbations lower MACE on all six benchmarks relative to this setting. Prediction centering further improves five benchmarks, notably GoogleMap (1.88 to 1.62) and RGB-NIR (0.96 to 0.69), with a 0.01-pixel difference in the opposite direction on GoogleEarth (1.08 vs. 1.07). These results support the regularizing role of geometric diversity and the utility of centering it on the current estimate. A complementary held-out geometric consistency test in Appendix B.3 examines how residual perturbations affect the translator's spatial behavior during adaptation.

5 Conclusions

We presented *Tailor-Made for One* (TMO), a parameter-efficient framework that adapts a strong zero-shot prior to each unlabeled test pair. Prediction-centered data synthesis and iterative anchor refresh construct pair-specific supervision while updating only translator LoRA adapters. Synthetic pretraining provides accurate initial alignment, and adaptation consistently improves it across six benchmarks. Gains with alternative components further support the framework's modularity.

AI USE STATEMENT

Generative AI tools were used only for code checking, manuscript language refinement, and organizational suggestions. They were not involved in the development of the methodology, experimental design, hypothesis formulation, implementation of the proposed method, or interpretation of experimental results. All AI-assisted outputs were manually reviewed and verified by the authors. The authors take full responsibility for the final manuscript, including all text, claims, code, and other AI-assisted content.

REFERENCES

Honggyu An, Jin Hyeon Kim, Seonghoon Park, Jaewoo Jung, Jisang Han, Sunghwan Hong, and Seungryong Kim. Cross-View Completion Models Are Zero-Shot Correspondence Estimators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1103–1115, 2025.

Matthew Brown and Sabine Süsstrunk. Multi-Spectral SIFT for Scene Category Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 177–184, 2011.

Wenxiao Cai and Wankou Yang. Object-Level Geometric Structure Preserving for Natural Image Stitching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1926–1934, 2025.

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 (CVPR)*, pp. 1869–1878, 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 (CVPR)*, pp. 9833–9842, 2023.

Chen Chen, Kangcheng Bin, Ting Hu, Jiahao Qi, Xingyue Liu, Tianpeng Liu, Zhen Liu, Yongxiang Liu, and Ping Zhong. Fusion Meets Diverse Conditions: A High-diversity Benchmark and Baseline for UAV-based Multimodal Object Detection with Condition Cues. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27958–27967, 2025.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple Baselines for Image Restoration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 17–33, 2022.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep Image Homography Estimation. *arXiv preprint arXiv:1606.03798*, 2016.

Ding Ding, Yueming Pan, Ruoyu Feng, Qi Dai, Kai Qiu, Jianmin Bao, Chong Luo, and Zhenzhong Chen. HomoGen: Enhanced Video Inpainting via Homography Propagation and Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22953–22962, 2025.

Yuanbiao Gou, Haiyu Zhao, Boyun Li, Xinyan Xiao, and Xi Peng. Test-Time Degradation Adaptation for Open-Set Image Restoration. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235, pp. 16167–16177, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Zhaohong Huang, Yuxin Zhang, Wenjing Liu, Fei Chao, and Rongrong Ji. Prototype-Based Test-Time Adaptation of Vision-Language Models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

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 (CVPRW)*, pp. 4269–4278, 2021.

Till Kroeger, Radu Timofte, Dengxin Dai, and Luc Van Gool. Fast Optical Flow Using Dense Inverse Search. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 471–488, 2016.

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 (CVPR)*, pp. 7649–7658, 2020.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-Aware Dual Adversarial Learning and a Multi-Scenario Multi-Modality Benchmark to Fuse Infrared and Visible for Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5792–5801, 2022a.

Shuaicheng Liu, Yuhang Lu, Hai Jiang, Nianjin Ye, Chuan Wang, and Bing Zeng. Unsupervised Global and Local Homography Estimation with Motion Basis Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7885–7899, 2022b.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ty Nguyen, Steven W. Chen, Shreyas S. Shivakumar, Camillo Jose Taylor, and Vijay Kumar. Unsupervised Deep Homography: A Fast and Robust Homography Estimation Model. *IEEE Robotics and Automation Letters*, 3(3):2346–2353, 2018.

Hyoungseob Park, Anjali Gupta, and Alex Wong. Test-Time Adaptation for Depth Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20519–20529, 2024.

Jiangwei Ren, Xingyu Jiang, Zizhuo Li, Dingkang Liang, Xin Zhou, and Xiang Bai. MINIMA: Modality Invariant Image Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23059–23068, 2025.

Gernot Riegler, David Ferstl, Matthias Rüther, and Horst Bischof. A Deep Primal-Dual Network for Guided Depth Super-Resolution. *arXiv preprint arXiv:1607.08569*, 2016.

Ruizhi Shao, Gaochang Wu, Yuemei Zhou, Ying Fu, Lu Fang, and Yebin Liu. LocalTrans: A Multiscale Local Transformer Network for Cross-Resolution Homography Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14870–14879, 2021.

Xuelun Shen, Zhipeng Cai, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, and Cheng Wang. GIM: Learning Generalizable Image Matcher from Internet Videos. In *The Twelfth International Conference on Learning Representations*, 2024.

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.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 9229–9248, 2020.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Kunpeng Wang, Keke Chen, Chenglong Li, Zhengzheng Tu, and Bin Luo. Alignment-Free RGB-T Salient Object Detection: A Large-Scale Dataset and Progressive Correlation Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7780–7788, 2025.

Xiaolong Wang, Runsen Xu, Zhuofan Cui, Zeyu Wan, and Yu Zhang. Fine-Grained Cross-View Geo-Localization Using a Correlation-Aware Homography Estimator. In *Advances in Neural Information Processing Systems*, volume 36, pp. 5301–5319, 2023.

Xiao Wu, Zi-Han Cao, Ting-Zhu Huang, Liang-Jian Deng, Jocelyn Chanussot, and Gemine Vivone. Fully-Connected Transformer for Multi-Source Image Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2071–2088, 2025.

Baochen Xiong, Xiaoshan Yang, Yaguang Song, Yaowei Wang, and Changsheng Xu. Modality-Collaborative Test-Time Adaptation for Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26732–26741, 2024.

Jiacheng Ying, Hui-Liang Shen, and Si-Yuan Cao. Unaligned Hyperspectral Image Fusion via Registration and Interpolation Modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 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 (CVPR)*, 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 (CVPR)*, pp. 16685–16694, 2025.

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 (ECCV)*, 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 (CVPR)*, 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 (ECCV)*, 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 (CVPR)*, pp. 15950–15959, 2021.

Xuchang Zhong, Xu Cao, Jinke Feng, and Hao Fang. HOLO: Homography-Guided Pose Estimator Network for Fine-Grained Visual Localization on SD Maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41376–41385, 2026.

Shenglong Zhou, Zhiwei Xiong, and Feng Wu. Test-time adaptation via style and structure guidance for histological image registration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7677–7685, 2024.

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 (CVPR)*, pp. 25932–25941, 2024.

A Additional Experimental Details

A.1 Evaluation Protocol

Table 5 lists the complete evaluation splits (5,731 pairs), label sources, and resizing canvases.

Table 5: Evaluation data and label sources. Native denotes dataset-provided four-corner annotations; synthetic denotes a deterministic projective perturbation generated by our unified evaluator. All rerun methods use 128×128 patches.

Dataset	Modality pair	Pairs	Label	Canvas
OPT-SAR	Optical–SAR	2,000	Synthetic	192×192
GoogleMap	Aerial–map	888	Native	192×192
GoogleEarth	Cross-condition aerial	850	Native	192×192
DPDN	RGB–depth	1,000	Synthetic	192×192
RGB-NIR	Visible–NIR	153	Synthetic	256×256
M3FD	Visible–infrared	840	Synthetic	192×192

For synthetic labels, we follow the widely used four-point perturbation protocol (DeTone et al., 2016; Koguciuk et al., 2021): each corner-displacement coordinate is sampled independently and uniformly from the integers in $[-32, 32]$, with one fixed homography per pair. A frozen manifest fixes displacements and crop locations across methods. Images are bilinearly resized to the listed canvases. We then apply the homography warp to each source image before cropping the source and target into 128×128 patches. MACE is measured in patch coordinates.

Baseline evaluation and result sources. The supervised and unsupervised baselines are retrained under their respective training protocols and evaluated alongside TMO using identical resized inputs, pair directions, frozen label manifests, and pixel-coordinate scales. The zero-shot CCNet results are quoted from You et al. (2026) under its published evaluation protocol, as its code and pretrained weights were unavailable at the time of our experiments.

A.2 Conditional Translator Architecture

The translator uses the source image to determine spatial structure and the condition image to guide appearance, aiming to change modality-specific appearance while retaining the source geometry.

Source path. A four-level encoder with NAFBlock-style (Chen et al., 2022) gated residual blocks extracts source features through progressive downsampling. The first three levels provide decoder skips; the bottleneck uses two-dimensional rotary positional encoding. The encoder and skips depend only on the source, which provides the spatial reference.

Condition path. Multiscale non-overlapping patches provide local appearance descriptors with scale identities but no image-position embeddings. Differentiable intensity histograms and Fourier-magnitude statistics complement these descriptors with intensity and texture information. Learned-query attention pooling and position-free token mixing aggregate them into compact appearance tokens. The pooled tokens and image intensity means and standard deviations form a global style vector. This branch supplies appearance cues without passing a spatial feature map to the decoder.

Fusion and output. Eight bottleneck transformer blocks combine source self-attention with cross-attention to condition tokens. Three decoder stages progressively upsample features using pixel shuffle, fuse source skips, and inject appearance through gated cross-attention, with source features as queries and condition tokens as keys and values. NAFBlock-style residual blocks refine fused features, modulated by spatially constant channel-wise scales and shifts from the style vector. The output combines the source with a learned residual. The output head and style projections are initialized near zero for an initial near-identity mapping.

LoRA configuration and parameter counts. The translator backbone has 44.24M parameters. LoRA ($r = 8$, $\alpha = 8$, no dropout) covers all linear and non-grouped 1×1 convolutional layers, adding 1.39M trainable parameters (3.15% of the backbone). The homography estimator is excluded from both counts. At test time, LoRA adapters specialize the translator to each pair while the pretrained backbone remains frozen. The alternative U-Net in Table 3 follows SSHNet (Yu et al., 2025).

A.3 Pretraining and Adaptation Details

Data and geometric synthesis. Pretraining uses only MS COCO (Lin et al., 2014) `train2017`, with synthetic validation on `val2017`. Paired views share a random RGB crop, bilinearly resized to 128×128 for the translator or 192×192 for the estimator. Each corner-displacement coordinate is sampled independently and uniformly from the integers in $[-32, 32]$. Inverse warping uses bilinear interpolation and reflection padding, applied directly to translator crops or to estimator canvases before central 128×128 cropping. Estimator reference corners bound this central patch, leaving a 32-pixel context margin. In each translation direction, the reconstruction target shares the geometry of the corresponding source input. Training transforms are resampled online.

Appearance transformations. Generic estimator augmentation and domain-randomized translator condition corruption use 12 operation types: occlusion, noise, sharpening, solarization, glass blur, color jitter, channel jitter, color inversion, gradual intensity variation, blur, weather effects, and JPEG compression. Table 6 groups them into ten families and illustrates additional synthesis mechanisms: gamma remapping, edge responses, resolution changes, and quantization. Operation counts are sampled uniformly from one to four for the estimator and two to three for the domain-randomized translator branch. We then uniformly select that many distinct operations and compose them with random parameters. Estimator augmentation affects either or both views with independently sampled transformations. Translator corruption varies by appearance family and affects only condition images, leaving reconstruction targets unchanged. These operations introduce no global projective warp; glass blur permits bounded local displacements.

Table 6: Representative appearance transformations used for synthetic data generation, grouped by family with illustrative mappings. Sampling and usage details are given in the text. Outputs are clipped to $[0, 1]$.

Transformation family	**Representative mapping**
Channel mixing	$A\mathbf{x}(u) + \mathbf{b}_c$
Tone mapping	$a\mathbf{x} + b$, $\mathbf{x}^\gamma$, $1 - \mathbf{x}$, $\mathrm{Solarize}_\tau(\mathbf{x})$
Detail / edge response	$\mathbf{x} + \lambda(\mathbf{x} - K * \mathbf{x})$, $\frac{1}{2}\big(g + \mathcal{N}(\lVert\nabla g\rVert_2)\big)$
Blur / resampling	$K * \mathbf{x}$, $\mathrm{Up}_s(\mathrm{Down}_s(\mathbf{x}))$
Glass blur	$\mathrm{Blur}[\mathbf{x}(u + \boldsymbol{\delta}(u))]$
Noise	$\mathbf{x} + \boldsymbol{\epsilon}$, $\mathbf{x} \odot \mathbf{n}$
Spatial gain / bias	$\mathbf{a}(u) \odot \mathbf{x}(u) + \mathbf{b}(u)$
Masking	$M \odot \mathbf{x} + (1 - M) \odot \mathbf{c}$
Fog / rain	$(1 - \alpha) \odot \mathbf{x} + \alpha \odot \mathbf{v}$, $\mathbf{x} + \mathbf{r}$
Quantization / JPEG	$\mathrm{round}((q-1)\mathbf{x})/(q-1)$, $\mathrm{JPEG}_Q(\mathbf{x})$

$\mathbf{x}$ is a normalized image and u a pixel location. $A, \mathbf{b}_c$ specify pointwise channel mixing and bias; a, b, γ control contrast, brightness, and tone; λ controls sharpening. Solarization inverts intensities at or above threshold τ. K is a smoothing kernel, s the resize scale, and $\boldsymbol{\delta}(u)$ a bounded local displacement. g denotes grayscale intensity; $\mathcal{N}$ rescales per-pixel gradient magnitudes to $[0, 1]$, and grayscale responses are replicated across channels. $\mathbf{a}(u), \mathbf{b}(u)$ are smooth gain and bias fields; $\boldsymbol{\epsilon}, \mathbf{n}$ denote additive and multiplicative noise; $M, \mathbf{c}$ a binary retention mask and fill; $\alpha, \mathbf{v}, \mathbf{r}$ fog opacity, fog intensity, and rain streaks. q is the number of quantization levels and Q the JPEG quality. $*$ and $\odot$ denote convolution and elementwise multiplication.

Pretraining. The homography estimator and modality translator are each pretrained for 200,000 steps with a global batch size of 32. Both use AdamW (Loshchilov & Hutter, 2017) with weight decay 10^{-5} and OneCycle learning-rate schedules with a 5% warm-up phase. Peak learning rates are 2×10^{-4} for the estimator and 10^{-4} for the translator, with cosine and linear annealing, respectively. The translator loss weights are $\lambda_{\mathrm{reg}}^{\mathrm{pre}} = 4$ and $\lambda_{\mathrm{feat}}^{\mathrm{pre}} = 1.5$.

Test-time adaptation. The translator's rank-8 LoRA adapters are optimized for 300 steps per test pair with a batch size of 2. Adaptation uses AdamW with weight decay 10^{-5} and a fixed learning rate of 1.2×10^{-4}. The same adaptation settings are used across all six benchmarks. Reconstruction and feature losses have weights 1 and 1.5, respectively. Both pretrained backbones remain frozen, and adapters are reset between pairs. At every update, each residual corner-displacement coordinate is sampled independently and uniformly from the integers in $[-24, 24]$ ($\eta = 24$ pixels). Test-time warps use zero padding and recompute validity masks. Reconstruction excludes invalid pixels;

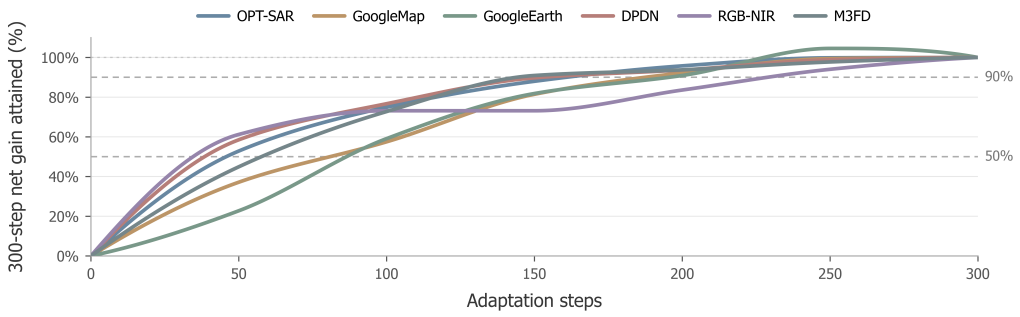

Figure 4: Fraction of each dataset's 300-step net MACE reduction attained during adaptation. Dashed lines mark 50% and 90%. Shape-preserving interpolation of 50-step measurements is used only for visualization.

feature masking follows Equation (12), with masked inputs and full-grid averaging without valid-area renormalization.

B Additional Experimental Results

B.1 Accuracy Gains across Adaptation Steps

We examine how quickly adaptation realizes the improvements reported in Table 2. Let $E_d(n)$ denote MACE on dataset d at step n. The percentage of its 300-step net gain attained is

$$g_d(n) = 100\frac{E_d(0) - E_d(n)}{E_d(0) - E_d(300)}. \tag{13}$$

Rapid empirical convergence. Figure 4 shows rapid early progress followed by diminishing returns. By step 100, all six datasets have attained more than 50% of their 300-step net MACE reduction. By step 200, five datasets exceed 90%, while RGB-NIR reaches 83.6%. The smaller remaining gains indicate that adaptation approaches a stable regime within the 300-step budget.

This analysis quantifies how the observed accuracy gains accumulate with the number of adaptation updates, using the 300-step result as a common reference. Values above 100% indicate that an earlier checkpoint achieves lower MACE than the 300-step endpoint.

B.2 Convergence Analysis Beyond MACE

Prior studies typically report final dataset-average MACE for homography estimation. Our framework introduces test-time adaptation into cross-modal homography estimation, motivating additional statistics to assess stability across adaptation iterations. Table 7 reports the median and first-to-third quartile interval (Q1–Q3) of per-pair average corner errors every 50 adaptation steps.

Progressive improvement and stable convergence beyond MACE. Median and quartile errors generally decrease as adaptation proceeds, showing that improvements extend beyond dataset-average MACE. By step 300, all medians and both quartiles are lower than at initialization on every dataset, and the interquartile width decreases by 12.6%–64.9%. For example, GoogleMap's Q1–Q3 shifts from 0.706–3.792 to 0.297–1.381 pixels. Between steps 250 and 300, median errors change by at most 0.027 pixels across all datasets, supporting stable late-stage convergence despite minor fluctuations.

B.3 Effect of Prediction-Centered Perturbations on Geometric Consistency

Purpose. The ablations in Table 4 show that prediction-centered perturbations reduce MACE, but do not directly establish whether they discourage the translator from absorbing residual misalignment. We therefore conduct a separate diagnostic to assess whether these perturbations help preserve the translator's geometric consistency.

Table 7: Evolution of per-pair average corner errors during test-time adaptation: median and interquartile interval (Q1–Q3), in pixels.

	OPT-SAR		GoogleMap		GoogleEarth	
Step	**Median**	**Q1–Q3**	**Median**	**Q1–Q3**	**Median**	**Q1–Q3**
0	11.820	6.885–19.148	1.557	0.706–3.792	0.795	0.542–1.326
50	7.708	4.555–14.989	1.165	0.531–2.671	0.783	0.536–1.284
100	6.138	3.961–12.454	1.021	0.438–2.449	0.750	0.523–1.232
150	5.432	3.625–10.611	0.786	0.372–1.906	0.739	0.516–1.222
200	5.086	3.439–9.419	0.680	0.315–1.544	0.735	0.515–1.215
250	4.829	3.308–9.112	0.610	0.300–1.428	0.724	0.526–1.211
300	4.802	3.292–9.107	0.597	0.297–1.381	0.729	0.520–1.205
	DPDN		**RGB-NIR**		**M3FD**	
Step	**Median**	**Q1–Q3**	**Median**	**Q1–Q3**	**Median**	**Q1–Q3**
0	6.476	3.935–11.147	0.222	0.155–0.425	2.588	1.706–6.189
50	5.174	3.151–9.077	0.197	0.132–0.325	2.407	1.608–4.665
100	4.692	3.012–8.206	0.205	0.122–0.317	2.287	1.581–3.761
150	4.328	2.802–7.351	0.194	0.121–0.301	2.231	1.599–3.535
200	4.106	2.614–6.992	0.189	0.124–0.299	2.165	1.554–3.360
250	3.886	2.396–6.650	0.194	0.127–0.300	2.105	1.516–3.193
300	3.863	2.440–6.578	0.192	0.127–0.291	2.102	1.533–3.154

Geometric consistency measure. We compare applying a source warp before translation with applying the same warp afterward, keeping the condition fixed. For the translator $\mathcal{G}_n$ after n adaptation steps, the two outputs are

$$\mathbf{U}_{n,H} = \mathcal{G}_n\big(\mathcal{W}(\mathbf{X}_s;\mathbf{H}),\mathbf{X}_t\big), \qquad \mathbf{V}_{n,H} = \mathcal{W}\big(\mathcal{G}_n(\mathbf{X}_s,\mathbf{X}_t);\mathbf{H}\big). \tag{14}$$

If translation preserves source geometry, these outputs should share the same spatial structure. Their residual misalignment therefore measures how consistently translation responds to the applied warp.

We estimate residual flow between the outputs using Dense Inverse Search (DIS) (Kroeger et al., 2016) and take its median magnitude over valid structural pixels. Smaller residuals indicate that less additional displacement is needed to align the outputs. Residuals are measured in pixels; lower is better. Because DIS can also respond to appearance changes, it provides evidence of geometric consistency rather than a guarantee of exact source-geometry preservation.

Protocol. We use the first 100 test pairs from each benchmark (600 pairs total), 128×128 patches, and 300 adaptation steps. Both variants retain prediction centering: *No residual* uses only the current alignment, whereas *Random residual* additionally samples each residual corner-displacement coordinate uniformly from the integers in $[-24, 24]$ at every update. These correspond to *No residual perturbation* and *Prediction-centered (default)* in Table 4. All other settings, initialization, and per-image seeds are matched, isolating the effect of geometric diversity around the anchor.

For each image, we average the median DIS residuals over fixed test warps to obtain one score. The warps are shared across variants and iterations; neither they nor their scores enter optimization or model selection. Table 8 reports dataset means and medians at step 300, while Table 9 tracks the equally weighted mean across the six datasets throughout adaptation.

Results and analysis. Table 8 shows that random residual perturbations reduce macro-average DIS by 69.1%, from 2.417 to 0.747 pixels. Both mean and median decrease on every dataset, with lower residuals on 96.5% of pairs overall. The lower medians and high paired improvement rate indicate that the benefit is broadly shared across samples and modality pairs, rather than being driven by a few large errors.

Table 9 shows that both variants start at 1.940 pixels. Without residual perturbations, inconsistency initially decreases but later exceeds its starting value; with perturbations, it decreases at every reported step to 0.747 pixels. This contrast suggests that geometric diversity helps counter spatial drift during adaptation. Together, these results support prediction-centered geometric diversity as an implicit equivariance regularizer that constrains spatial inconsistency during adaptation.

Table 8: DIS residuals at step 300 on 100 pairs per dataset (pixels; lower is better). Each image score averages the median residuals from three fixed test warps. Mean and median summarize these scores. Both variants retain prediction centering. Pairs improved denotes the percentage with lower scores under random residual perturbations.

	No residual		Random residual		Pairs
Dataset	Mean	Median	Mean	Median	**improved**
OPT-SAR	1.591	1.365	**0.482**	**0.387**	96%
GoogleMap	2.219	1.484	**0.457**	**0.324**	98%
GoogleEarth	1.939	0.604	**0.398**	**0.139**	97%
DPDN	3.499	2.711	**1.723**	**1.378**	94%
RGB-NIR	1.318	0.295	**0.185**	**0.093**	98%
M3FD	3.933	3.123	**1.238**	**0.783**	96%
Macro	2.417	–	**0.747**	–	96.5%

Table 9: Evolution of DIS residuals on 600 pairs (pixels; lower is better). Image scores average over three fixed test warps; reported residuals are equally weighted six-dataset means. Mean reduction and pairs improved compare random residuals against no residuals at each step.

	DIS residual		Mean	Pairs
Step	No residual	Random residual	**reduction**	**improved**
0	1.940	1.940	0.0%	–
50	1.729	**1.023**	40.8%	80.5%
100	1.999	**0.890**	55.5%	90.2%
150	2.274	**0.866**	61.9%	92.3%
200	2.419	**0.812**	66.4%	95.3%
250	2.448	**0.764**	68.8%	95.7%
300	2.417	**0.747**	69.1%	96.5%

C Discussion and Limitations

Adaptation cost and budget. TMO requires additional computation for adaptation on each test pair. We use a shared optimization schedule and a 300-step budget across all benchmarks, without dataset-specific tuning or label-based stopping. Under this setting, all six datasets recover over half of their 300-step MACE reduction by step 100, and five exceed 90% by step 200 (Appendix B.1). These results suggest that shorter runs can retain substantial improvements when computation is limited. The appropriate budget may still depend on pair difficulty, motivating label-free stopping as a direction for improving efficiency.

Geometric initialization. TMO uses the current estimate to define its local adaptation neighborhood. Severely inaccurate anchors or limited overlap may weaken self-supervision, making recovery through local refinement difficult. Our synthetic pretraining already provides strong zero-shot priors with RHWF, IHN, and MCNet, and TTA further improves all six benchmarks with each estimator (Table 3). Our contribution centers on TTA for homography estimation, complementing advances in pretrained models. Future estimators may provide more accurate anchors, and the gains across current architectures suggest that adaptation could remain beneficial with stronger initialization.