Few-Shot LoRA Adaptation of a Flow-Matching Foundation Model for Cross-Spectral Object Detection

Maxim Clouser Kia Khezeli John Kalantari

Yrikka Inc.

max@yrikka.com, kia@yrikka.com, john@yrikka.com

Abstract

Foundation models for vision are predominantly trained on RGB data, while many safety-critical applications rely on non-visible modalities such as infrared (IR) and synthetic aperture radar (SAR). We study whether a single flow-matching foundation model pre-trained primarily on RGB images can be repurposed as a cross-spectral translator using only a few co-measured examples, and whether the resulting synthetic data can enhance downstream detection. Starting from FLUX.1 Kontext, we insert low-rank adaptation (LoRA) modules and fine-tune them on just 100 paired images per domain for two settings: RGB→IR on the KAIST dataset and RGB→SAR on the M4-SAR dataset. The adapted model translates RGB images into pixel-aligned IR/SAR, enabling us to reuse existing bounding boxes and train object detection models purely in the target modality. Across a grid of LoRA hyperparameters, we find that LPIPS computed on only 50 held-out pairs is a strong proxy for downstream performance: lower LPIPS consistently predicts higher mAP for YOLOv11n on both IR and SAR, and for DETR on KAIST IR test data. Using the best LPIPS-selected LoRA adapter, synthetic IR from external RGB datasets (LLVIP, FLIR ADAS) improves KAIST IR pedestrian detection, and synthetic SAR significantly boosts infrastructure detection on M4-SAR when combined with limited real SAR. Our results suggest that few-shot LoRA adaptation of flow-matching foundation models is a promising path toward foundation-style support for non-visible modalities.

1. Introduction

The rapid progress of foundation models in computer vision has been largely confined to the visible spectrum, where massive datasets of RGB images underpin training of general-purpose models like CLIP and Stable Diffusion [23, 25]. By contrast, many real-world applications in infrared (IR) and synthetic aperture radar (SAR) domains lack comparable data scale and pre-trained models [4]. This gap is critical: autonomous driving, surveillance, and remote sensing all require perception beyond visible light, yet current vision foundation models struggle to generalize to modalities such as IR or SAR. The motivation for this work arises from the need to extend foundation models beyond the visible spectrum, leveraging their powerful learned representations to benefit low-resource sensing domains.

One promising strategy is cross-spectral image translation, using models trained on abundant visible imagery to synthesize corresponding IR or SAR views. Prior GAN-based translators can produce realistic outputs but require dataset-specific training and can be unstable across spectra, sometimes hallucinating structures or distorting IR brightness [14, 22, 26, 28, 30]. Diffusion-based translation can be more stable [9, 11], but training a diffusion model from scratch for each sensor pair remains costly [17]. We therefore ask whether a single pre-trained foundation generator can be repurposed into a reusable cross-spectral translator with only a small number of co-measured examples. Concretely, we adapt FLUX.1 Kontext [8, 18], a latent rectified-flow transformer pre-trained on large-scale RGB data, by inserting Low-Rank Adaptation (LoRA) modules [10] and fine-tuning only these parameters on a small paired set of aligned RGB–IR or RGB–SAR examples. This yields a parameter-efficient mapping that preserves the base model prior while keeping data and compute requirements low.

We validate our approach on two datasets: KAIST multispectral (RGB–IR pedestrian scenes) [13] and M4-SAR (an RGB–SAR satellite imagery benchmark for object detection) [2]. With as few as 100 paired examples per domain, LoRA-adapted FLUX.1 Kontext generates realistic IR or SAR images that are useful for downstream detection. We find that LPIPS (Learned Perceptual Image Patch Similarity) scores between synthetic and real images [29] strongly correlates with target-domain mAP: lower LPIPS on a small validation set reliably predicts better detector performance. Using the best LPIPS-selected LoRA adapter, we show two practical uses of cross-spectral augmentation:

(1) translating RGB-only LLVIP and FLIR ADAS images into the KAIST IR domain, which improves KAIST pedestrian detection over training on limited real IR alone; and (2) augmenting M4-SAR with synthetic SAR generated from co-registered RGB images, yielding a notable mAP gain over using only real SAR. In summary, our contributions include the following:

- **Parameter-efficient cross-spectral translation.** We adapt a single flow-matching diffusion foundation model for cross-spectral image translation (RGB→IR, RGB→SAR) via LoRA fine-tuning. With only 100 paired examples per dataset, the resulting adapters produce high-fidelity translations and act as reusable cross-spectral translators, illustrating an effective way to extend vision foundation models beyond the visible spectrum with modest data and compute.
- **Correlation of perceptual quality with detection.** We show that LPIPS perceptual similarity [29] on a small validation set is a reliable indicator of downstream utility: lower LPIPS correlates with higher detection mAP on both IR and SAR tasks, supporting its use as a proxy metric when labeled detection data are scarce.
- **Boosting detection with cross-spectral data augmentation.** We translate additional RGB datasets into the target modality and reuse their labels to augment detector training, improving performance in low-data IR and SAR settings.

2. Related Work

2.1. Cross-Spectral Image Translation

Early approaches to cross-spectral translation employed generative adversarial networks. Pix2pix demonstrated supervised translation using paired images and a conditional GAN objective [14]. To relax the need for pairing, CycleGAN introduced cycle-consistency losses enabling translation between unpaired datasets [30]. These frameworks inspired numerous extensions: UNIT and MUNIT incorporated stochastic mappings for multimodal outputs [12, 19], and many domain-specific GAN models have been proposed for spectral translation. For RGB→IR translation, specialized GANs such as ThermalGAN [16] are used to generate thermal images for person re-identification, and InfraGAN [24] is used to improve realism of IR outputs. IR translation has also been studied with attention to stability and detail. For example, Ma et al. fuse multi-scale features in a pix2pix-based IR generator, and other works introduce architectural variants to better translate thermal face images to visible, all within GAN-style frameworks [16, 22, 24, 26, 28].

In remote sensing, RGB–SAR translation techniques based on CycleGAN are used to compensate for sensing gaps, e.g., generating RGB-like images from SAR when clouds obscure satellite imagery [27]. Variants of CycleGAN and related architectures introduce additional structural constraints such as segmentation-guided losses to maintain object shapes, road topology, or coastline structure during SAR→RGB translation [22, 26–28]. Despite these advances, GAN-based cross-spectral translation remains challenging. Typical failure modes include misalignment of fine details, brightness distortions, and hallucinated structures in the generated images [22, 26, 28].

Modern diffusion models offer a compelling alternative due to their stability and sample quality [9]. Denoising diffusion and score-based models have been applied to cross-spectral tasks. For instance, VI-Diff employs a diffusion model for unpaired RGB→IR translation in person re-identification [11]. VI-Diff reports improved fidelity of synthetic IR for re-identification, albeit with high computational cost and dataset-specific training [11, 17]. Physics-Informed Diffusion (PID) builds on HADAR-style physics-based IR formation [1] and introduces a TeV decomposition together with physics-informed reconstruction and TeV-space consistency losses for physically grounded IR image translation [20]. Our work differs in that we do not train a diffusion model from scratch for each spectral pair. Instead, we fine-tune a general pre-trained model. By leveraging a strong diffusion model trained on large-scale RGB data [8, 18, 25], we obtain excellent cross-spectral translation with only a fraction of the data and training time that a bespoke GAN or diffusion would require.

2.2. LoRA and Parameter-Efficient Fine-Tuning of Generative Models

Low-Rank Adaptation (LoRA) [10] injects small trainable low-rank matrices into a pre-trained network, enabling parameter-efficient fine-tuning without modifying most of the original weights. Originally proposed for adapting large language models, LoRA has since been widely used to customize diffusion-based image generators such as Stable Diffusion to new styles or concepts with modest compute [10, 17, 25]. We follow this paradigm and adapt the FLUX.1 Kontext model to IR and SAR translation using LoRA adapters that comprise less than 1% of the base model parameters, yet suffice to imprint the cross-spectral mapping in data-scarce regimes.

2.3. Flow-Matching and Score-Based Generative Models in Vision

Score-based diffusion models such as DDPMs generate images via iterative denoising and have revolutionized image synthesis [9]. Flow matching generalizes this family by training continuous normalizing flows to match a time-indexed probability flow between noise and data, reproducing diffusion behavior while improving stability [18]. Rectified flow transformers further scale these ideas to high-

resolution image generation with competitive quality and efficient sampling [3]. We leverage FLUX.1 Kontext, a latent rectified-flow model that supports flexible conditioning and in-context image editing by accepting both an input image and a text prompt [5, 8]. Its unified training on generation and editing tasks [5, 8] makes FLUX.1 a natural backbone for cross-spectral translation.

2.4. Synthetic IR/SAR Data for Object Detection

Prior work has used synthetic images to improve IR and SAR object detection when real annotations are scarce. For infrared pedestrian detection, several works use CycleGAN-style RGB→IR translation while reusing RGB bounding boxes, improving IR detectors without extra thermal labels [22, 26, 28]. In SAR, augmentation by translation is less explored, but M4-SAR shows that combining RGB and SAR inputs improves detection over SAR alone, suggesting that RGB imagery provides complementary structure [2]. Our work follows this line by generating synthetic IR and SAR with a foundation model and using them as additional training data. Unlike prior GAN-based approaches, we leverage a single LoRA-adapted flow-matching model rather than training task-specific translators from scratch.

3. Method: LoRA-Adapted Flow Matching for Cross-Spectral Translation

3.1. Problem Setting

We consider cross-spectral translation between a source modality x^s (RGB) and a target modality x^t (IR or SAR). Given a small set of aligned pairs

$$\mathcal{D}_{\text{pair}} = \{(x_i^s, x_i^t)\}_{i=1}^N,$$

and a larger set of labeled target-domain images $\mathcal{D}_{\text{det}}$ with bounding boxes $\mathcal{B}$, our goal is to:

1. Learn a conditional generator G_θ that translates x^s into a synthetic target image $\hat{x}^t = G_\theta(x^s)$ that is pixel-aligned with x^s.
2. Use synthetic images $\hat{x}^t$ to train object detectors that operate purely in the target domain (IR or SAR), either by augmenting $\mathcal{D}_{\text{det}}$ with synthetic data or by training on synthetic target images alone.

The main constraint is that $|\mathcal{D}_{\text{pair}}|$ is very small (100 pairs per dataset in our experiments) and is used exclusively to train the LoRA adapters, reflecting the scarcity of co-measured cross-spectral data. Figure 1 summarizes the overall pipeline from LoRA adaptation to detector training. We denote the small paired subset used for LoRA training as the *Sensor Sample* split, a disjoint paired subset for validation as *Sensor Val*, and the full detection training images as the *Train* split.

3.2. Base Model: FLUX.1 Kontext

We build on FLUX.1 Kontext, a 12B-parameter rectified-flow transformer trained in the latent space of an autoencoder. The model unifies image generation and editing via flow matching: given a time $t \in [0, 1]$, a latent state z_t, and conditioning c, the network predicts the probability flow $v_\phi(z_t, t, c)$ that transports a simple base distribution (e.g., Gaussian noise) to the data distribution. Training minimizes a squared-error objective between the predicted and ground-truth flow along a chosen interpolation path.

FLUX.1 Kontext supports in-context editing. The conditioning c can include both a text prompt and a reference image. We exploit this capability by providing the source-domain image x^s as the conditioning image and an instruction as the text conditioning to render the same scene in the target modality. In all experiments we use a fixed dataset-specific prompt, e.g., "Convert this to an IR image from the KAIST sensor" for KAIST and "Convert this to a SAR image" for M4-SAR, shared across all training pairs. We deliberately avoid image-specific captions; a detailed per-image textual description is an interesting direction for future work.

3.3. LoRA-Adapted Flow Matching

Directly fine-tuning FLUX.1 Kontext for each new spectral pair would require updating billions of parameters and substantial compute. Instead, we adopt Low-Rank Adaptation (LoRA), inserting small trainable matrices into selected layers while freezing the base model.

For a weight matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ in the attention or MLP projections of the transformer backbone, LoRA introduces a low-rank increment:

$$W' = W + \Delta W, \quad \Delta W = \frac{\alpha}{r} AB,$$

where $A \in \mathbb{R}^{d_{\text{out}} \times r}$, $B \in \mathbb{R}^{r \times d_{\text{in}}}$ are trainable, $r \ll \min(d_{\text{out}}, d_{\text{in}})$ is the rank, and α is a scaling factor. Only A and B are updated; all original weights in FLUX.1 Kontext remain fixed. We attach these LoRA adapters to the query, key, value, and output projections in each self-attention layer, as well as to the linear projections in the MLP sub-blocks of the image transformer backbone, yielding additional trainable parameters on the order of $\sim 1\%$ of the base model while remaining expressive enough to capture the cross-spectral mapping in each domain.

In our experiments, we define a "training step" as a single optimizer update. We fine-tune LoRA using a batch size of 1. We adopt the standard LoRA initialization in which the initial effective weight update is $\Delta W = 0$ [10]. Specifically, we initialize the down-projection with Kaiming uniform and the up-projection to zeros [6, 21], ensuring training starts from the base model's behavior. Investigating alternative LoRA initialization schemes, and their effects in low-data regimes, is a promising direction for future work.

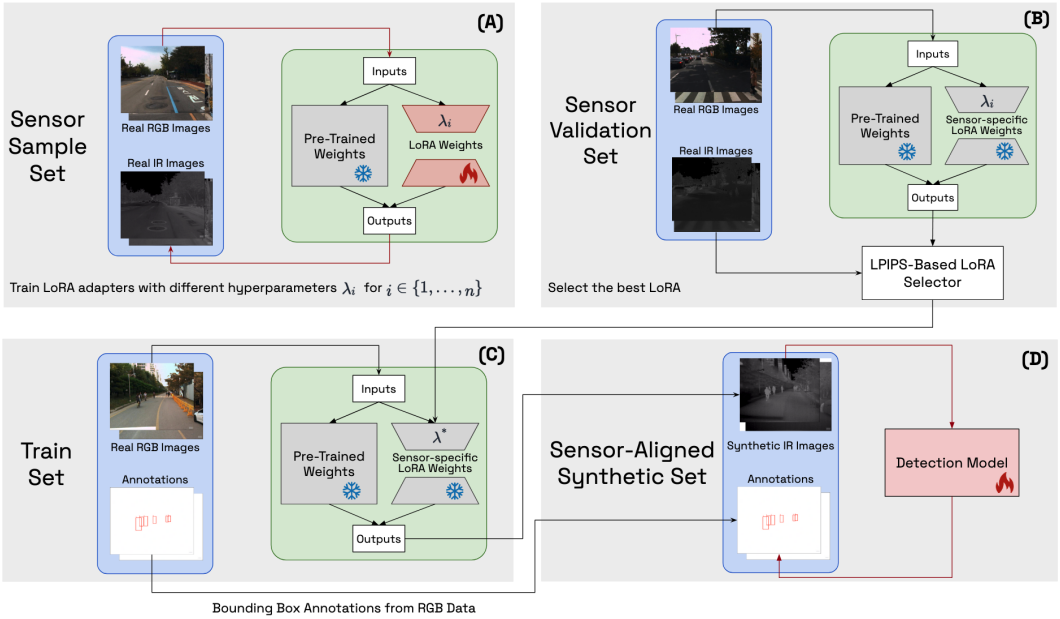

Figure 1. Overview of our pipeline for LoRA-adapted flow-matching cross-spectral translation and detection. (A) A small *Sensor Sample* split of paired RGB and IR/SAR images is used to train multiple LoRA configurations on top of a frozen FLUX.1 Kontext base model. (B) A separate *Sensor Val* split is translated and scored against real images with LPIPS to select the best LoRA. (C) The selected LoRA λ^* is applied to the RGB *Train* split to generate a sensor-aligned synthetic target-domain set, reusing the original RGB bounding box annotations. (D) The sensor-aligned synthetic set, optionally combined with real target-domain images, is used to train an object detection model in the target modality. All panels show KAIST RGB→IR; the same pipeline is applied for RGB→SAR.

3.4. Model Selection via LPIPS

We use LPIPS (Learned Perceptual Image Patch Similarity) [29] as a model-selection metric. LPIPS measures the ℓ_2 distance between deep feature representations of a synthetic image and its real counterpart, computed from a fixed AlexNet backbone, and correlates strongly with human perceptual judgments of image similarity. Lower LPIPS indicates closer perceptual match to the target modality. We specifically selected LPIPS over alternatives such as PSNR, SSIM, and Fréchet Inception Distance (FID) for two reasons. First, LPIPS has been shown to correlate more strongly with human perceptual similarity than PSNR or SSIM [29]. This makes LPIPS more suitable for our cross-spectral translation setting, where preserving semantic structure and textures is more important than minimizing pixel-wise error. Second, we operate in an extremely low-data regime on the sensor-specific validation splits (50 paired images per domain), where distribution-level metrics such as FID become statistically fragile.

For each dataset (KAIST and M4-SAR), we sweep a grid of LoRA hyperparameters. Let λ denote a LoRA hyperparameter configuration (learning rate, rank r with $\alpha = r$, and number of training steps), and let $\{\lambda_1, \ldots, \lambda_n\}$ be the set of configurations in the sweep (with $n = 15$ in our experiments):

- learning rate $\in \{1 \times 10^{-4}, 5 \times 10^{-4}\}$,
- rank $r \in \{16, 32\}$ with $\alpha = r$,
- training steps $\in \{1\text{k}, 3\text{k}, 6\text{k}, 10\text{k}, 30\text{k}, 40\text{k}\}$.

For 1k, 3k, and 6k steps we train all 2×2 combinations of learning rate and rank r (with $\alpha = r$), while for 10k, 30k, and 40k we instantiate only the configuration with learning rate 5×10^{-4} and rank $r = 16$ (i.e., $\alpha = 16$), resulting in $n = 15$ LoRA configurations $\{\lambda_i\}_{i=1}^{n}$ per dataset, chosen to fit within our compute and time constraints. Each configuration λ_i is trained on the *Sensor Sample* split (100 paired images). We then:

1. Translate the 50-image *Sensor Val* split using each LoRA adapter.
2. Compute LPIPS between each synthetic image and its real IR/SAR counterpart.
3. Use the average LPIPS as a cross-spectral validation score.

The configuration with lowest LPIPS is selected as the best LoRA for that dataset, which we denote by λ^*. This avoids training detectors for every configuration and turns LoRA selection into a computationally efficient generative evaluation problem.

3.5. Synthetic Dataset Construction for Detection

Once the best LoRA is chosen:

- For KAIST and M4-SAR, we translate every image in the *Train* split from the source modality (RGB) into the target (IR/SAR), obtaining a synthetic training set $\hat{\mathcal{D}}_{\text{Train}}$.
- Because translation is pixel-aligned, we reuse the original bounding boxes from the real *Train* split without modification.
- For the cross-dataset KAIST experiments, we similarly translate RGB-only images from LLVIP and FLIR ADAS into the KAIST IR domain and reuse their labels.

Object detectors are then trained on real-only, synthetic-only, or real + synthetic combinations, depending on the experiment.

4. Experimental Setup

4.1. Datasets and Splits

4.1.1. KAIST Multispectral Pedestrian

KAIST contains 95k RGB–IR images captured from a vehicle-mounted beam splitter, with annotations for person, people, and cyclist classes [13]. We focus on the *person* class and restrict our experiments to daytime sequences to avoid dark RGB frames that may induce hallucinated structures during translation. We define five non-overlapping splits:

- **Sensor Sample** (100 pairs): unlabeled, pixel-aligned RGB–IR pairs used exclusively for LoRA training.
- **Sensor Val** (50 pairs): unlabeled RGB–IR pairs used for LPIPS-based LoRA selection.
- **Train** (800 pairs): pixel-aligned RGB–IR pairs with bounding boxes used for detector training.
- **Val** (200 images): IR-only frames with annotations used for model selection.
- **Test** (911 images): IR-only frames with annotations used for final evaluation.

These splits represent a small, curated subset of the full KAIST corpus (which contains 95k RGB–IR images), chosen to emulate a data-scarce setting. KAIST is organized into numbered driving sequences; we use sequence 1 exclusively for *Sensor Sample* and *Sensor Val*, sequences 0 and 2 for *Train* and *Val*, and sequences 9–11 for *Test*, ensuring no frame overlap across splits.

4.1.2. External RGB Datasets: LLVIP and FLIR ADAS

For cross-dataset scaling, we leverage two additional multispectral pedestrian datasets but use only their RGB views during training:

- **LLVIP**: 30,976 images (15,488 RGB–thermal pairs) from a binocular RGB–IR sensor in low-light conditions, with pedestrian bounding boxes [15].
- **FLIR ADAS**: 26,000 RGB–thermal pairs at 640×512 resolution, with more than 520k bounding boxes over 15 categories (person, bicycle, car, bus, etc.) [7].

We discard their IR channels, translate the RGB images into the KAIST IR domain using the best KAIST LoRA, and reuse the original labels. The resulting synthetic frames are appended to the KAIST *Train* split to test whether external RGB corpora can be reused for IR detection via cross-spectral translation.

4.1.3. M4-SAR RGB–SAR Benchmark

M4-SAR provides co-registered RGB–SAR image pairs at 10 m (VH) and 60 m (VV) resolution. We use only the 10 m VH subset (files 1–56087) and restrict detection to two classes: *bridge* and *harbor*.

From the official train and test partitions we construct:

- **Sensor Sample** (100 pairs): unlabeled, pixel-aligned RGB–SAR pairs used exclusively for LoRA training.
- **Sensor Val** (50 pairs): unlabeled, pixel-aligned RGB–SAR pairs used for LPIPS-based LoRA selection.
- **Train** (1600 pairs): pixel-aligned RGB–SAR pairs with bounding boxes used for detector training.
- **Val** (400 images): SAR-only frames with annotations used for model selection.
- **Test** (200 images): SAR-only frames with annotations used for final evaluation.

The dataset is highly imbalanced: bridges account for roughly 94–96% of all annotations across splits, with harbors making up the remainder. Similar to KAIST, our dataset partitions form a modest subset of the full M4-SAR dataset, reflecting a practical regime where high-resolution SAR annotations are scarce.

4.2. LoRA Training and Synthetic Generation

For each combination of hyperparameters in our grid, we fine-tune LoRA modules on the *Sensor Sample* split of the corresponding dataset using the flow-matching loss inherited from FLUX.1 Kontext and condition on the source image and a fixed dataset-specific instruction prompt described in Section 3.2. After training:

1. We translate the 50 *Sensor Val* pairs and compute mean LPIPS against real IR/SAR images.
2. We translate the *Train* split (800 KAIST pairs / 1600 M4-SAR pairs) to obtain synthetic IR/SAR training sets.

For all detection experiments, we train and evaluate detectors exclusively on target-modality images (IR or SAR) and their ground-truth bounding boxes. RGB images are used only as inputs to the translation model to generate synthetic target-modality data. All synthetic images are pixel-aligned with their corresponding source RGB frames.

4.3. Object Detection Models

We study two object detection families:

- **YOLOv11n** (Ultralytics): a modern one-stage detector optimized for efficiency. We train for 30 epochs with batch size 16 and default Ultralytics hyperparameters.

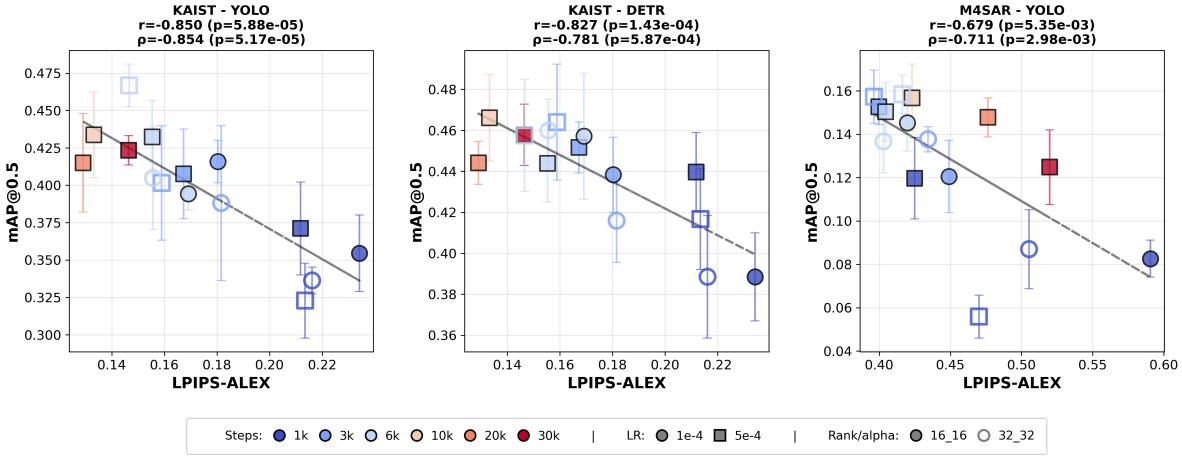

Figure 2. LPIPS on *Sensor Val* versus YOLOv11n / DETR mAP@0.50 on the real *Test* sets. From left to right: (i) KAIST with YOLOv11n, (ii) KAIST with DETR, and (iii) M4-SAR with YOLOv11n. Each point corresponds to a LoRA configuration: color encodes the number of LoRA training steps, marker shape encodes learning rate, and filled versus unfilled markers encode LoRA rank. Points show the mean over 5 runs per configuration and error bars indicate ± 1 standard deviation. Solid lines show least-squares linear fits. Panel titles report Pearson (r) and Spearman (ρ) correlation coefficients with associated p-values, all indicating strong negative correlations between LPIPS and downstream detection performance.

- **DETR**: a transformer-based detector representing a conceptually different architecture. We train for 30 epochs with learning rate 1×10^{-5} and batch size 8.

Unless otherwise noted, detectors are trained on either real or synthetic versions of the *Train* split and evaluated on real *Test* images only. For each training configuration, we report mean and standard deviation of mAP over five independent train and test runs using different random seeds.

5. Results

5.1. LPIPS as a Proxy for Downstream Detection

For each dataset and LoRA hyperparameter configuration, we:

1. Compute the average LPIPS on *Sensor Val* between synthetic and real images.
2. Train YOLOv11n (and DETR for KAIST) on the corresponding synthetic *Train* set.
3. Evaluate mAP on the real *Test* split.

Figure 2 visualizes LPIPS versus mAP@0.50 across all LoRA configurations for KAIST (YOLOv11n and DETR) and M4-SAR (YOLOv11n).

For KAIST and YOLOv11n (left panel of Fig. 2), each point corresponds to a LoRA configuration. We observe a clear negative correlation: models with lower LPIPS on *Sensor Val* achieve higher mAP@0.50 on KAIST *Test*. The linear fit captures this trend quantitatively, with Pearson correlation $r = -0.85$ ($p = 5.88 \times 10^{-5}$) and Spearman correlation $\rho = -0.85$ ($p = 5.17 \times 10^{-5}$).

The middle panel of Fig. 2 shows the same behavior for DETR on KAIST, again with a strong negative relationship between LPIPS and downstream mAP (Pearson $r = -0.83$, Spearman $\rho = -0.78$, both highly significant).

In the right panel of Fig. 2, the same pattern arises in the RGB-to-SAR setting on M4-SAR: lower LPIPS corresponds to higher detection mAP on the *Test* set, despite class imbalance and domain complexity (Pearson $r = -0.68$ ($p = 5.35 \times 10^{-3}$), Spearman $\rho = -0.71$ ($p = 2.98 \times 10^{-3}$)).

Across both datasets and architectures, low LPIPS on just 50 validation pairs reliably predicts which LoRA configuration yields the best downstream detection performance. Practically, this means we can select a LoRA adapter without training detectors for all configurations, drastically reducing search cost.

Figure 3 illustrates how these quantitative trends manifest visually. For both KAIST and M4-SAR, the best LoRA (ranked by YOLOv11n mAP@0.50) produces IR/SAR images whose global contrast and local structures more closely match the real sensors, particularly around pedestrians, bridges, and harbor structures. The worst-performing LoRA, in contrast, exhibits blurrier backgrounds, distorted object shapes, and spurious textures, which likely reduce the utility of these images for detector training.

5.2. Cross-Dataset Extension on KAIST

We next ask whether the best KAIST LoRA can be used to translate external RGB corpora into KAIST-style IR and thereby extend the effective IR training set. Using the best KAIST adapter (chosen via LPIPS), we translate 400 FLIR ADAS RGB frames and 500 LLVIP RGB frames into synthetic KAIST-like IR images and append them to the real

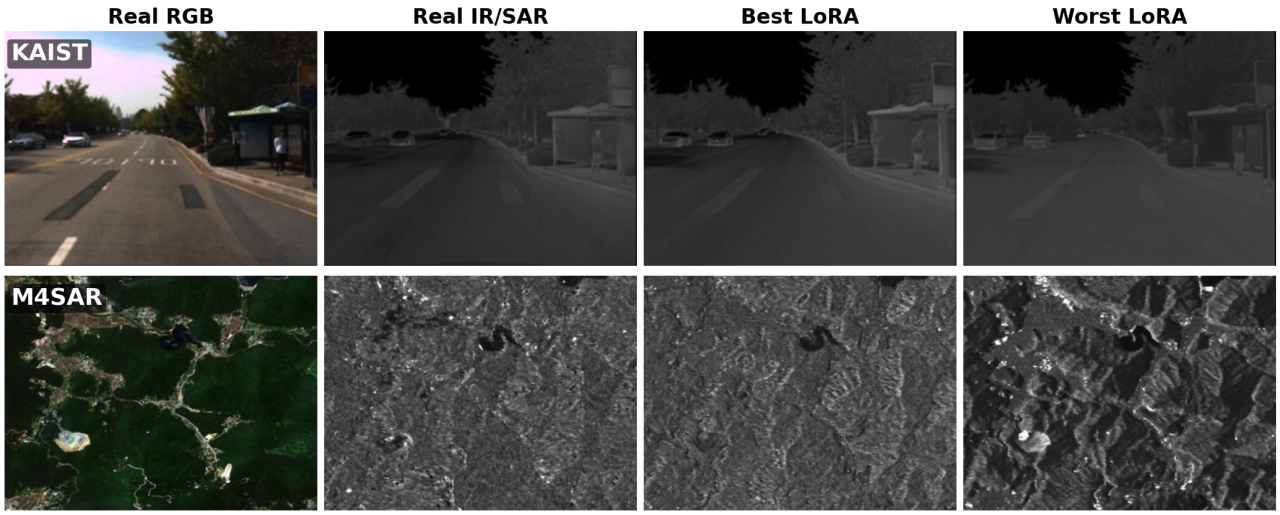

Figure 3. Examples of cross-modal image translation on the KAIST (RGB–IR) and M4-SAR (RGB–SAR) datasets. For each dataset and scene, columns show the input RGB image, the corresponding real IR/SAR image, and synthetic IR/SAR images generated by the best- and worst-performing LoRA configurations, ordered from left to right. Best and worst are ranked by downstream YOLOv11n mAP@0.50.

Table 1. mAP on the KAIST *Test* set for YOLOv11n and DETR with and without synthetic IR translated from external RGB datasets.

	YOLOv11n		**DETR**	
Train Set	@0.50	@[0.5:0.95]	@0.50	@[0.5:0.95]
Real KAIST (baseline)	0.50 ± 0.02	0.22 ± 0.01	0.48 ± 0.02	0.19 ± 0.01
+ FLIR-synth (400 imgs)	$\mathbf{0.54 \pm 0.02}$	0.23 ± 0.01	0.47 ± 0.02	0.18 ± 0.01
+ LLVIP-synth (500 imgs)	$\mathbf{0.54 \pm 0.02}$	0.22 ± 0.01	0.48 ± 0.01	0.19 ± 0.00

KAIST *Train* set. Table 1 reports results:

- For YOLOv11n, mAP@0.50 improves from 0.50 (KAIST-only baseline) to 0.54 with FLIR-synth and 0.54 with LLVIP-synth.
- For DETR, performance remains essentially unchanged or slightly degraded.

These gains are particularly notable because no additional IR annotations are required: we simply reuse labels from FLIR and LLVIP after translation. Qualitatively, synthetic IR images preserve pedestrian shapes and coarse scene layout while adapting contrast and background clutter to match KAIST's thermal domain.

Taken together, these experiments show that a single LoRA-adapted flow-matching model can act as a practical translator that unlocks RGB-only datasets for IR detection. In terms of perceptual quality, the LPIPS of our best KAIST LoRA (0.129 on *Sensor Val* using only 100 aligned RGB–IR training pairs in *Sensor Sample*) is on par with the strongest Physics-Informed Diffusion (PID) configuration on KAIST (0.128 LPIPS), and sits near the bottom of the 0.37–0.13 LPIPS range reported for GAN and diffusion baselines in Table 1 of Mao et al. [20].

Table 2. M4-SAR: YOLOv11n detection performance with real and synthetic SAR.

Train Set	**# Real**	**# Synth**	**mAP@0.50**	**mAP@ [0.50:0.95]**
Real-only	1600	0	0.19 ± 0.01	0.06 ± 0.01
Synthetic-only	0	5000	0.18 ± 0.02	0.06 ± 0.01
Real + Synthetic	1600	5000	$\mathbf{0.25 \pm 0.01}$	$\mathbf{0.09 \pm 0.01}$

5.3. Scaling SAR Detection with Synthetic SAR

Finally, we investigate whether synthetic SAR generated from RGB images can boost detection performance on M4-SAR. Using the best M4-SAR LoRA (again chosen by *Sensor Val* LPIPS), we translate 5000 RGB images into synthetic SAR and train YOLOv11n in three regimes:

1. Real-only: 1600 real SAR images.
2. Synthetic-only: 5000 synthetic SAR images.
3. Real + Synthetic: 1600 real + 5000 synthetic.

Table 2 summarizes the results. Real-only provides the baseline, synthetic-only is slightly worse, and combining real and synthetic SAR yields a substantial boost in both

mAP@0.50 and mAP@[0.50:0.95] (a >30% relative gain over the real-only baseline at mAP@0.50).

Despite severe class imbalance and the complex statistics of SAR imagery, synthetic SAR clearly acts as an effective data augmenter when combined with limited real SAR.

6. Discussion

6.1. LPIPS as a Practical Selection Signal

Our experiments support a simple but powerful observation: LPIPS on a tiny validation set is a strong proxy for downstream object detection performance. LPIPS is computed on only 50 paired images per dataset, yet it predicts trends in mAP obtained from training full detectors on hundreds or thousands of translated images. This has practical implications:

- LoRA selection can be guided by generative quality alone, avoiding expensive end-to-end detector retraining for each hyperparameter setting.
- In low-resource settings where detection labels are scarce (or available only for a subset of the domain), LPIPS offers an inexpensive surrogate for the task relevance of synthetic data.

Nevertheless, LPIPS is an image-level metric. It does not explicitly account for object-level fidelity or radiometric correctness, which may become important for fine-grained scientific applications.

6.2. Synthetic Data as a Bridge Across Modalities

Our results on KAIST and M4-SAR highlight complementary roles for synthetic data:

- RGB→IR translation enables cross-dataset expansion. One can harvest RGB-only datasets and map them into the IR domain, extending training distributions without new IR sensors or annotations.
- RGB→SAR translation provides within-dataset scaling. Synthetic SAR increases sample diversity for underrepresented classes and viewpoints, boosting detector performance when combined with real SAR.

Importantly, synthetic-only training lags behind real-only baselines in the SAR case, underscoring that current generators exhibit a *sim-to-real gap* and do not yet fully replace real measurements. Instead, they act as amplifiers for scarce real data.

6.3. Architectural Sensitivity to Synthetic Data

YOLOv11n consistently benefits from synthetic augmentation, especially on KAIST, whereas DETR shows minimal gains and occasionally slight regressions. Several factors may contribute:

- One-stage detectors such as YOLOv11n may better exploit the improved background diversity and object variety introduced by synthetic data.
- DETR's transformer architecture can be more data intensive and may require larger or more diverse datasets to realize benefits from augmentation.
- Synthetic artifacts (for example, subtle texture inconsistencies) might interact differently with each model's inductive biases.

Exploring detector architectures tailored to synthetic-heavy regimes, or joint training of generator and detector, is an interesting direction for future work.

6.4. Limitations and Future Directions

We summarize limitations and directions for future work:

- **Paired data requirement.** While we use only 100 co-registered pairs per modality, some settings may lack any paired data. A promising direction is to fine-tune a text-to-image backbone with LoRA using unpaired text supervision, then reuse the same adapters for image-to-image translation, reducing or eliminating the need for co-registered pairs.
- **Limited modalities and tasks.** We focus on RGB→IR and RGB→SAR for pedestrian and infrastructure detection. Other modalities such as radio-frequency (RF) imagery and tasks such as segmentation, tracking, and change detection remain to be explored. We also operate in a low-data regime (hundreds of training images), which may underestimate the gains achievable when combining foundation-model translation with large-scale labeled IR/SAR datasets.
- **Radiometric fidelity.** Our method optimizes for perceptual similarity rather than physical accuracy. For scientific remote sensing applications, enforcing sensor-specific radiometric constraints or leveraging physics-informed priors may be necessary.
- **Single foundation model.** All experiments use FLUX.1 Kontext as the base generator. Investigating other foundation models (e.g., text–vision or remote sensing FMs) and comparing adaptation strategies could reveal when flow-matching is most advantageous.
- **Detector scope.** We evaluate one lightweight one-stage detector (YOLOv11n) and one transformer-based detector (DETR). Extending this evaluation to additional architectures (including real-time and few-shot detectors) under the same data and compute constraints is an important direction for future work.

Despite these limitations, our results demonstrate a promising path forward: a single flow-matching foundation model, adapted via lightweight LoRA modules, can serve as a reusable cross-spectral translator that meaningfully improves IR and SAR detection in low-data regimes.

References

[1] Fanglin Bao, Xueji Wang, Shree Hari Sureshbabu, Gautam Sreekumar, Liping Yang, Vaneet Aggarwal, Vishnu N Boddeti, and Zubin Jacob. Heat-assisted detection and ranging. *Nature*, 619(7971):743–748, 2023. 2

[2] Wen Chao et al. M4-SAR: A benchmark for multi-modal multi-resolution SAR-optical detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3

[3] Xi Chen et al. Rectified flow: Enhancing flow-based generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[4] Saikat Chowdhury et al. Foundation models beyond the visible spectrum. *arXiv preprint arXiv:2310.04135*, 2023. 1

[5] Patrick Esser et al. Decoupled contextual transformer for image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[6] Hugging Face. Peft developer guide: Lora. https://huggingface.co/docs/peft/en/developer_guides/lora. Accessed: 2026-01-06. 3

[7] FLIR Systems Inc. Flir thermal dataset for algorithm training. https://www.flir.com/oem/adas/adas-dataset-form/, 2018. Accessed: 2025-08-27. 5

[8] FLUX Team. FLUX.1 kontext technical report. Technical Report, Hugging Face, 2025. 1, 2, 3

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2

[10] Edward J. Hu et al. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 1, 2, 3

[11] Han Huang et al. VI-Diff: Visible-to-infrared image translation via diffusion models. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2023. 1, 2

[12] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[13] Soonmin Hwang, Jaesik Park, Namil Kim, Youngchan Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baselines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 5

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[15] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, Shengjie Liu, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3496–3504, 2021. 5

[16] Vadim V. Kniaz, Vladimir A. Knyaz, Jörg Hladuvka, Walter G. Kropatsch, and Konrad Schindler. ThermalGAN: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018. 2

[17] Yiming Li et al. Diffusion models in vision: A survey. *arXiv preprint arXiv:2307.10458*, 2023. 1, 2

[18] Yaron Lipman et al. Flow matching for generative modeling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 1, 2

[19] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[20] Fangyuan Mao, Jilin Mei, Shun Lu, Fuyang Liu, Liang Chen, Fangzhou Zhao, and Yu Hu. Pid: physics-informed diffusion model for infrared image generation. *Pattern Recognition*, 169:111816, 2026. 2, 7

[21] Ostris. Lora variants. https://deepwiki.com/ostris/ai-toolkit/9.2-lora-variants. Accessed: 2026-01-06. 3

[22] Rui Qiu et al. Thermal infrared image synthesis from RGB image using CycleGAN. *International Journal of Computer Applications*, 2020. 1, 2, 3

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. 1

[24] Yifei Ren et al. InfraGAN: Thermal image synthesis via conditional generative adversarial network for person re-identification. *Sensors*, 22(13), 2022. 2

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[26] Yan Song et al. Bridging the domain gap for thermal infrared pedestrian detection using GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3

[27] Zhen Wang, Hong Zhang, Feng Xu, and Ya-Qiu Jin. SAR-to-optical image translation based on CycleGAN. *Remote Sensing*, 12(3), 2020. 2

[28] Jiabin Zhang et al. Improved CycleGAN for thermal image translation. *Sensors*, 21(6), 2021. 1, 2, 3

[29] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 4

[30] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2