ANYVIA: A UNIVERSAL VISIBLE-INFRARED ADAPTER FOR UAV PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

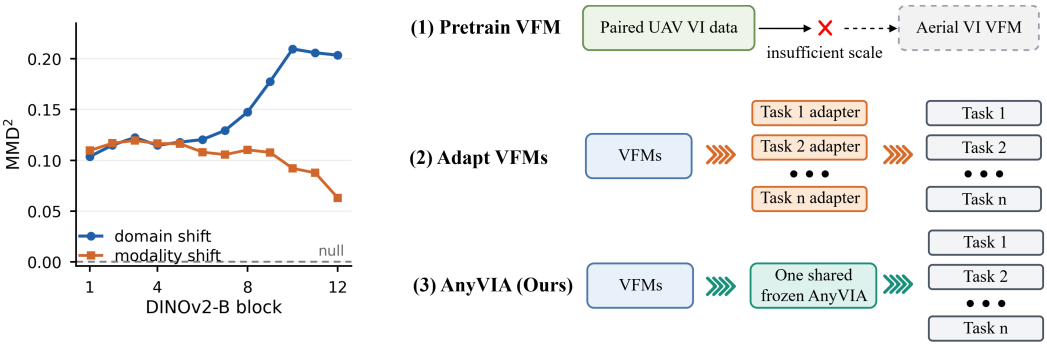
Vision foundation models (VFMs) provide transferable representations. However, aerial visible-infrared (VI) perception benefits from jointly modeling the complementary cues present in visible and infrared imagery. Most VFMs are pretrained primarily on RGB data and are not explicitly optimized for such cross-spectral modeling. Pretraining a dedicated aerial VI VFM is constrained by the cost of collecting large numbers of cross-spectrally registered image pairs, while existing VI adapters are typically tied to particular encoders and tasks. We introduce AnyVIA, a 299K-parameter encoder-agnostic adapter that transforms and spatially upsamples frozen VFM features for aerial VI perception. AnyVIA is pretrained once and then reused in frozen form across heterogeneous VFMs and downstream tasks. For each modality, a Raw Guidance Encoder (RGE) derives attention queries and keys solely from the input image, while host features enter only as values after parameter-free channel folding. This design decouples AnyVIA's learned weights from the host feature width. During pretraining, shared box prompts anchor both modalities to the same objects, while a directional consistency objective aligns refined feature directions with those of frozen host features extracted from corresponding higher-resolution crops. To provide dense supervision for cross-spectral representation learning, we further construct UAVRGBT235K, a pretraining resource for unmanned aerial vehicle (UAV) VI adaptation. It comprises 117,668 image pairs and 7.15 million automatically generated instance masks, substantially expanding the object-level annotations available for aerial VI pretraining. Across eight held-out UAV benchmarks spanning five perception tasks, the frozen AnyVIA yields improvements for diverse encoders. These results suggest that AnyVIA offers a scalable and reusable alternative to pretraining a dedicated VFM for aerial VI perception.

1 INTRODUCTION

Visible-infrared (VI) perception from unmanned aerial vehicles (UAVs) has been studied across a range of aerial tasks, including object detection (Sun et al., 2022b), semantic segmentation (Lee et al., 2024), salient object detection (Song et al., 2023), camouflaged object detection (Liu et al., 2026), object tracking (Zhang et al., 2022), and crowd counting (Peng et al., 2020). Task-specific UAV VI methods have made substantial progress in these settings, but their specialized designs limit generalization to new tasks and scenes. Vision foundation models (VFMs) (Radford et al., 2021; Oquab et al., 2023; Tschannen et al., 2025) offer a general-purpose alternative because they provide transferable features across visual tasks.

Aerial VI perception benefits from complementary visible and infrared cues, yet most generic VFMs are pretrained predominantly on RGB natural-image data and are not designed to model cross-spectral differences. The discrepancy between frozen DINOv2-B features extracted from natural and aerial VI images grows with depth (Fig. 1(a)), while visible and infrared representations remain distributionally distinct throughout the encoder. A dedicated aerial VI VFM could address both shifts, but training such a model from scratch requires large-scale paired data and substantial computation. Paired UAV VI imagery is costly to acquire because it requires calibrated dual-sensor platforms and cross-spectral registration, leaving existing public datasets modest in scale relative to modern VFM pretraining data. A more scalable alternative is to retain existing VFMs and learn a

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.



(a) Domain and modality shifts **(b) Three paradigms for aerial VI perception**

Figure 1: **Motivation for AnyVIA.** (a) Block-wise MMD^2 between frozen RGB-pretrained DINOv2-B features from 1,500 natural images and 1,500 aerial VI pairs (domain shift), and between the visible and infrared streams of the same aerial pairs (modality shift). The dashed line denotes a same-domain null obtained by splitting either domain into two halves. The increasing domain shift in deeper blocks motivates adaptation to aerial VI imagery. (b) Three paradigms for aerial VI perception. Existing paired UAV VI datasets remain modest in scale relative to modern VFM pretraining data, while task-specific VI adaptation typically learns an adapter for each downstream task. AnyVIA is pretrained once and reused, frozen, across VFMs and downstream tasks.

compact VI feature transformation. However, existing VI adapters are optimized for a particular encoder and downstream objective. Changing either typically requires retraining (Fig. 1(b)).

We introduce AnyVIA, a 299K-parameter encoder-agnostic VI adapter that jointly transforms and spatially upsamples frozen VFM features. AnyVIA is pretrained once and then reused in frozen form across encoders and tasks. For each modality, a Raw Guidance Encoder (RGE) derives attention queries and keys from the raw image, while the host features enter the transformation only as values. Fixed-width image guidance and parameter-free channel folding decouple the learnable weights from the host feature dimension, allowing the same adapter to process heterogeneous encoders. The two modality branches are jointly pretrained through shared-prompt mask supervision, which couples their predictions around common objects, and directional consistency, which aligns the refined feature directions with frozen host features extracted from corresponding higher-resolution crops. We pretrain AnyVIA on UAVRGBT235K, an aerial VI pretraining resource that we construct from 117,668 image pairs drawn from public datasets. We enrich their sparse source annotations with 7.15M instance masks generated by SAM2 (Ravi et al., 2025), substantially expanding the dense object-level supervision available for aerial cross-spectral adaptation.

Our contributions are summarized as follows: (i) We introduce AnyVIA, an encoder-agnostic VI feature adapter that transforms and spatially upsamples frozen VFM features and can be reused in frozen form across encoders and tasks. (ii) Fixed-width raw guidance and parameter-free channel folding keep its learned weights independent of the host feature width, while shared-prompt mask supervision and directional consistency jointly pretrain its two modality branches. (iii) We construct UAVRGBT235K, an aerial VI pretraining resource containing 117,668 public image pairs and 7.15M SAM2-generated instance masks. (iv) Across eight UAV benchmarks disjoint from the pretraining data and spanning five tasks, we demonstrate that a single frozen AnyVIA transfers across diverse encoders.

2 Related Work

2.1 Foundation Models for Aerial and Spectral Perception

Remote-sensing foundation models such as RingMo (Sun et al., 2022a) and SkySense (Guo et al., 2024) learn transferable representations from large-scale remote-sensing data, while RingMo-Aerial (Diao et al., 2025) specifically targets the characteristics of aerial imagery. Models for spectral perception follow several different designs. M-SpecGene (Zhou et al., 2025) learns RGBT rep-

resentations through Siamese masked autoencoding. MMPedestron (Zhang et al., 2024) jointly trains a unified encoder over dynamic combinations of RGB, infrared, depth, LiDAR, and event inputs. InfMAE (Liu et al., 2024) performs self-supervised pretraining on infrared imagery, whereas UNIP (Zhang et al., 2025a) targets infrared perception through attention-pattern distillation on mixed RGB–infrared data. These approaches strengthen the encoder itself. AnyVIA instead retains host encoders and learns a compact VI feature transformation that can be reused in frozen form.

2.2 Visible–Infrared Adaptation and Fusion

VFMs provide transferable representations for recognition and dense prediction. Representative families include vision–language models such as CLIP (Radford et al., 2021) and SigLIP2 (Tschannen et al., 2025), self-supervised encoders such as DINOv2 (Oquab et al., 2023) and DINOv3 (Simeoni et al., 2025), and promptable segmentation models such as SAM (Kirillov et al., 2023), SAM2 (Ravi et al., 2025), and SAM3 (Carion et al., 2026). These predominantly RGB-pretrained representations do not explicitly model visible–infrared spectral differences. UniRGB-IR (Yuan et al., 2025) augments a frozen RGB-pretrained VFM through adapter tuning, while StitchFusion (Li et al., 2025) propagates information between pretrained modality encoders during encoding. CMX (Zhang et al., 2023a) performs cross-modal feature rectification and fusion, and C2Former (Yuan & Wei, 2024) learns calibrated cross-modal interactions for RGB-IR detection. These methods optimize fusion or adaptation components for a selected encoder and downstream setting. AnyVIA instead pretrains one compact feature adapter and keeps it frozen across host encoders and tasks.

2.3 Encoder-Agnostic Feature Upsampling

Feature upsampling provides a closely related line of research. JAFAR (Couairon et al., 2025) reconstructs features at arbitrary target resolutions, while AnyUp (Wimmer et al., 2026), NAF (Chambon et al., 2026), and RaysUp (Ding et al., 2026) investigate upsampling across heterogeneous feature representations. These methods primarily reconstruct spatially denser feature maps from a single image stream. AnyVIA jointly transforms and spatially upsamples visible and infrared features for aerial VI adaptation. Its VI-specific pretraining enables one frozen adapter to transfer across host encoders and downstream tasks.

3 Method

3.1 Overview

Our goal is to learn VI adaptation once and reuse it across frozen vision encoders and UAV tasks. AnyVIA combines an encoder-agnostic architecture with VI pretraining (Fig. 2). A separate RGE learns spatial aggregation weights from each modality's raw image, independently of the host encoder's feature channels. The two branches are jointly pretrained on UAVRGBT235K by decoding their features under shared object prompts and supervising the sum of their mask predictions. A directional consistency loss aligns each modality's refined feature directions with those of frozen host features from the same image region. The pretrained adapter is then reused across frozen encoders and UAV tasks.

3.2 UAVRGBT235K

We construct UAVRGBT235K as an aerial VI pretraining resource by standardizing ten public UAV VI sources into $117{,}668$ image pairs. To expand their sparse source annotations, mostly a handful of boxes per image, into dense object-level supervision shared across modalities, we annotate each pair using frozen SAM2 (Ravi et al., 2025). Two branches run on every image, one prompted by the source boxes and one unprompted. The two mask sets are merged within each modality, and the resulting visible and infrared mask annotations are combined in the visible-image coordinate frame. UAVRGBT235K contains $7{,}154{,}516$ instance masks, 60.8 per pair. App. A gives the per-source composition and the annotation pipeline.

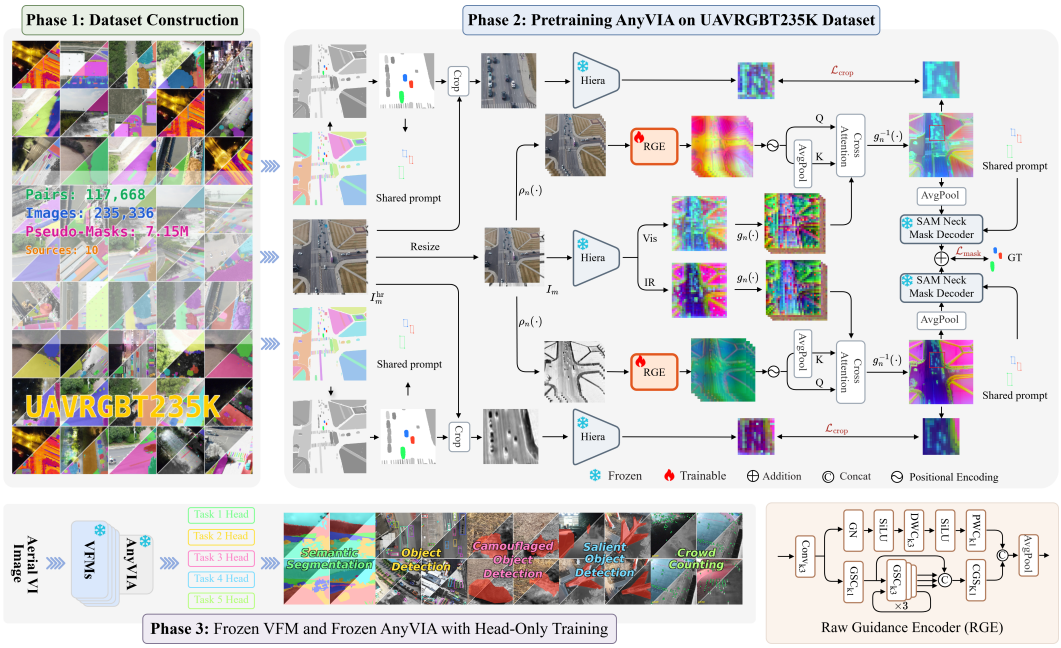


Figure 2: **Overview of AnyVIA. Phase 1** assembles ten public UAV VI sources into UAVRGBT235K and enriches their sparse annotations with shared instance masks generated by frozen SAM2. **Phase 2** jointly pretrains the two RGEs using SAM2's frozen Hiera-B+ trunk as the host encoder. Each RGE derives queries and pooled keys from its modality's image, while grouped encoder features provide the values. CrossAtten refines these features, which are average pooled and decoded by SAM2's frozen neck and mask decoder under a shared box prompt. The two modalities' mask logits are summed and supervised with $\mathcal{L}_{\text{mask}}$. $\mathcal{L}_{\text{crop}}$ compares refined feature windows with frozen host features from corresponding higher-resolution crops of the same modality via cosine similarity. **Phase 3** reuses the same pretrained AnyVIA across VFMs and UAV tasks.

3.3 AnyVIA

AnyVIA refines visible and infrared features using spatial aggregation weights derived from each modality's raw image. For modality $m \in \{v, r\}$, let $I_m \in \mathbb{R}^{3\times H\times W}$ be the resized image fed to a frozen encoder ϕ, and let $F_m = \phi(I_m) \in \mathbb{R}^{C\times h\times w}$ denote its spatial feature map. For each modality, an RGE supplies the attention queries and keys, while F_m provides the values without a learned projection. The resulting features $\hat{F}_m \in \mathbb{R}^{C\times 2h\times 2w}$ retain the host's channel dimension at a finer spatial resolution.

The two RGEs use the same architecture with independent parameter sets. For group width D, we use $n = \lceil C/D \rceil$ channel groups. For image guidance, I_m is replicated along the batch dimension by the parameter-free operator $\rho_n(\cdot)$, yielding $\bar{I}_m$ with one copy per group.

Within each RGE, all copies are processed by the same weights, starting with a stem $Z_m = \text{Conv}_{3\times3}(\bar{I}_m)$ that maps each copy to D channels. The stem output feeds the detail and context paths (Fig. 2).

The detail path extracts local spatial cues through depthwise filtering and pointwise channel mixing:

$$B_m^{\text{detail}} = \big(\text{PWConv}_{1\times1} \circ \text{SiLU} \circ \text{DWConv}_{3\times3} \circ \text{SiLU} \circ \text{GN}\big)(Z_m). \tag{1}$$

Here $\circ$ denotes function composition, and DWConv and PWConv denote depthwise and pointwise convolutions, respectively (Chollet, 2017). The final pointwise convolution projects the detail features from D to $D/2$ channels.

The context path gathers information over progressively larger neighborhoods through successive convolutions. The initial projection $U_m^0 = \text{GSC}_{1\times1}(Z_m)$ reduces the channel width to $D/2$. The

cascade $U_m^j = \mathrm{GSC}_{3\times3}^j(U_m^{j-1})$, $j = 1, 2, 3$, maintains this width. These features are aggregated as

$$B_m^{\text{context}} = \mathrm{CGS}_{1\times1}\big(U_m^0 \| U_m^1 \| U_m^2 \| U_m^3\big)\,, \tag{2}$$

where GSC and CGS denote GN–SiLU–Conv and Conv–GN–SiLU, respectively, and $\|$ denotes channel concatenation. The output projection maps the concatenated $2D$ channels to $D/2$. We concatenate the two $D/2$-channel outputs and average-pool the result to a spatial resolution of $2h \times 2w$. To incorporate relative spatial relationships into content-based attention, we apply 2D rotary positional encoding (Heo et al., 2024) to obtain the guidance field:

$$G_m = \mathrm{RoPE}\big(\mathrm{AvgPool}_{2h,2w}\big(B_m^{\text{context}} \| B_m^{\text{detail}}\big)\big) \in \mathbb{R}^{n\times D\times 2h\times 2w}. \tag{3}$$

A parameter-free fold $g_n(\cdot)$ partitions F_m into n channel groups, zero-pads the last group to width D, and stacks them along the batch dimension to obtain $g_n(F_m) \in \mathbb{R}^{n\times D\times h\times w}$. Each group is paired with its guidance copy. All copies within a modality supply identical queries at resolution $2h \times 2w$ and pooled keys at $h \times w$, yielding

$$\hat{F}_m = g_n^{-1}\big(\mathrm{CrossAtten}\big(G_m, \mathrm{AvgPool}_{h,w}(G_m), g_n(F_m)\big)\big)\,. \tag{4}$$

Here $\mathrm{CrossAtten}$ performs cross-scale neighborhood aggregation using neighborhood attention (Hassani et al., 2023), and $g_n^{-1}(\cdot)$ restores the original channel order and removes padding.

AnyVIA's learnable parameters reside in its RGEs, which generate image guidance of fixed width D. Parameter-free repetition, grouping, and restoration accommodate the host width C, so the same RGE weights can be applied to encoders with different feature dimensions. For downstream evaluation, the visible and infrared outputs are fused by parameter-free elementwise addition, $F_{\mathrm{VI}} = \hat{F}_v + \hat{F}_r$.

3.4 Pretraining pipeline

We jointly pretrain the two RGEs on UAVRGBT235K with shared-prompt mask supervision and a directional consistency loss. We use SAM2, whose Hiera-B+ trunk serves as the host encoder. The trunk, feature-pyramid neck, prompt encoder, and mask decoder remain frozen.

For each sampled instance mask Y, we derive a box prompt b and condition both modality branches on the same prompt. Average pooling returns the refined features to the host feature resolution, and the frozen neck and mask-decoding path $\mathcal{D}$ produce per-modality mask logits $\hat{M}_m$. We sum these predictions before applying mask supervision:

$$\hat{M}_m = \mathcal{D}\Big(\mathrm{AvgPool}_{h,w}(\hat{F}_m), b\Big)\,, \qquad \hat{M} = \hat{M}_v + \hat{M}_r. \tag{5}$$

We use the original SAM2 loss (Ravi et al., 2025), written as

$$\mathcal{L}_{\text{mask}} = \mathcal{L}_{\text{SAM2}}(\hat{M}, Y). \tag{6}$$

Supervising their summed logits couples the two RGEs through a common object target, while each RGE learns guidance from its own modality.

We further constrain feature directions within each modality using the frozen host as a reference. For each image pair, the sampled instance masks identify a foreground location around which we select a fixed-size window Ω, shared by both modalities. We extract the corresponding crop from I_m^{hr}, the higher-resolution counterpart of I_m, and encode it with ϕ. The resulting reference map and the window of $\hat{F}_m$ cover the same image region at the same spatial sampling density. Writing the refined feature window as $S_{m,\Omega} = \hat{F}_m[\Omega]$ and its reference as $T_{m,\Omega} = \phi(\mathrm{crop}_\Omega(I_m^{\text{hr}}))$, we minimize the directional consistency loss

$$\mathcal{L}_{\text{crop}} = \mathbb{E}_{m,\Omega,p}[1 - \cos(S_{m,\Omega}[p], T_{m,\Omega}[p])]\,, \tag{7}$$

where p indexes positions within Ω and cosine similarity is computed over channels. This loss aligns feature directions independently of their magnitudes. The total training loss is

$$\mathcal{L} = \mathcal{L}_{\text{mask}} + 2.8\,\mathcal{L}_{\text{crop}}. \tag{8}$$

Implementation details are provided in App. B.

4 EXPERIMENTS

We evaluate the transfer of a single AnyVIA across five UAV perception tasks and encoders with heterogeneous pretraining, compare it with task-specific VI adaptation, and analyze the effects of output resolution and input modality.

4.1 EXPERIMENTAL SETUP

4.1.1 DATASETS

We evaluate on eight public UAV VI benchmarks, all disjoint from UAVRGBT235K. CART (Lee et al., 2024) and Kust4K (Ouyang et al., 2025) provide semantic segmentation; UAV RGB-T 2400 (Song et al., 2023) and VIAC (Liu et al., 2026) cover salient and camouflaged object detection, respectively. DroneRGBT (Peng et al., 2020) is used for crowd counting. KUSTDrone (Wang et al., 2026) and UAV-CB (Huang et al., 2026) use horizontal bounding boxes, while ATR-UMOD (Chen et al., 2025) uses oriented bounding boxes.

4.1.2 EVALUATION METRICS

We report mIoU for semantic segmentation, S_α, F_β^w, MAE, and E_ϕ^{adp} for salient and camouflaged object detection, MAE and RMSE for crowd counting, and mAP for bounding-box detection.

4.1.3 IMPLEMENTATION DETAILS

Base and +AnyVIA share preprocessing, additive VI fusion, task heads, and training schedules. The encoder and AnyVIA remain frozen throughout downstream evaluation. We train all downstream models for 20 epochs with a batch size of 16. AdamW learning rates are 2×10^{-4} for mask prediction and counting, and 10^{-4} for detection. For detection, we also use an exponential moving average of model weights. Inputs are 512×512 for mask prediction, 512×640 for counting, and 448×448 for detection.

4.2 TRANSFER ACROSS DOWNSTREAM TASKS

Table 1: **Semantic segmentation performance on CART and Kust4K.** The metric is mIoU (%).

Dataset	Setting	SigLIP2-B	CLIP ViT-B	CLIP RN50	DINOv2-B	Franca-B	DINOv2-R-B
CART	Base	52.71	50.00	31.39	57.22	59.22	58.71
	+AnyVIA	58.10 (+5.39)	51.18 (+1.18)	34.30 (+2.91)	61.43 (+4.21)	61.52 (+2.30)	59.27 (+0.56)
Kust4K	Base	46.46	37.19	35.88	55.34	53.86	53.93
	+AnyVIA	52.22 (+5.76)	41.36 (+4.17)	38.74 (+2.86)	58.16 (+2.82)	56.21 (+2.35)	55.69 (+1.76)

4.2.1 SEMANTIC SEGMENTATION

For semantic segmentation, plain ViTs use a 1×1 classifier on the final feature map. For hierarchical encoders, we fold the last three feature maps to 256 channels without learned parameters, align and sum them, and then apply a 3×3 convolution, ReLU, and a 1×1 classifier. Tab. 1 reports AnyVIA with SigLIP2-B (Tschannen et al., 2025), CLIP ViT-B and CLIP RN50 (Radford et al., 2021), DINOv2-B (Oquab et al., 2023), Franca-B (Venkataramanan et al., 2026), and DINOv2-R-B (Darcet et al., 2024). AnyVIA improves mIoU for all six backbones on both CART and Kust4K. SigLIP2-B shows the largest gains, 5.39 and 5.76 points, respectively. With the multilevel head, CLIP RN50 gains 2.91 and 2.86 points; among plain ViTs, DINOv2-B gains 4.21 on CART and CLIP ViT-B gains 4.17 on Kust4K.

4.2.2 SALIENT OBJECT DETECTION

For salient object detection on UAV RGB-T 2400, the task head projects each selected feature level to 64 channels, aligns and concatenates the maps, and applies a 1×1 fusion layer, two

Table 2: **Salient object detection performance on UAV RGB-T 2400.**

Metric	Setting	CLIP ViT-B	SigLIP2-B	DINOv2-B	DINOv2-R-B	DINOv3-B	Franca-B	Swin-B
$S_\alpha \uparrow$	Base	69.40	77.31	80.22	80.22	80.42	79.82	81.63
	+AnyVIA	74.62 (+5.22)	80.83 (+3.52)	82.82 (+2.60)	82.69 (+2.47)	81.69 (+1.27)	81.49 (+1.67)	82.79 (+1.16)
$F_\beta^w \uparrow$	Base	34.85	49.08	54.74	54.92	55.90	53.59	56.61
	+AnyVIA	44.08 (+9.23)	57.10 (+8.02)	61.17 (+6.43)	59.49 (+4.57)	57.73 (+1.83)	55.92 (+2.33)	60.16 (+3.55)
MAE$\downarrow$	Base	2.61	1.81	1.41	1.44	1.34	1.47	1.57
	+AnyVIA	2.08 (-0.53)	1.44 (-0.37)	1.20 (-0.21)	1.20 (-0.24)	1.27 (-0.07)	1.29 (-0.18)	1.40 (-0.17)
$E_\phi^{\text{adp}} \uparrow$	Base	65.04	72.59	77.31	77.89	77.09	76.55	79.83
	+AnyVIA	70.86 (+5.82)	80.09 (+7.50)	82.34 (+5.03)	82.82 (+4.93)	80.23 (+3.14)	80.78 (+4.23)	82.00 (+2.17)

depthwise-separable convolutions, and a binary predictor. It reads the final map of plain encoders or the last three levels of hierarchical encoders. Tab. 2 reports seven backbones, including DINOv3-B (Simeoni et al., 2025) and Swin-B (Liu et al., 2021). AnyVIA improves all four metrics for every backbone in the table. CLIP ViT-B shows the largest gains in S_α ($+5.22$) and F_β^w ($+9.23$), together with the largest MAE decrease (from 2.61 to 2.08). SigLIP2-B has the largest E_ϕ^{adp} increase ($+7.50$).

4.2.3 Camouflaged object detection

For VIAC, we use the same task head as for UAV RGB-T 2400. Plain encoders provide their final feature map; hierarchical encoders provide their last three stages. Tab. 3 compares CLIP ViT-B, DINOv2-B, and Swin-B. AnyVIA improves all four metrics for each backbone. The F_β^w gains are 1.81, 2.08, and 2.05 points, respectively. CLIP ViT-B also gains 1.45 points in E_ϕ^{adp} and reduces MAE by 0.27.

Table 3: **Camouflaged object detection on VIAC.**

Metric	Setting	CLIP ViT-B	DINOv2-B	Swin-B
$S_\alpha \uparrow$	Base	86.79	90.89	90.13
	+AnyVIA	87.66 (+0.87)	90.98 (+0.09)	90.58 (+0.45)
$F_\beta^w \uparrow$	Base	72.59	80.27	79.15
	+AnyVIA	74.40 (+1.81)	82.35 (+2.08)	81.20 (+2.05)
MAE$\downarrow$	Base	2.90	1.78	2.01
	+AnyVIA	2.63 (-0.27)	1.70 (-0.08)	1.90 (-0.11)
$E_\phi^{\text{adp}} \uparrow$	Base	85.75	89.92	89.64
	+AnyVIA	87.20 (+1.45)	90.33 (+0.41)	89.82 (+0.18)

4.2.4 Object detection

For object detection, we use RF-DETR-S (Robinson et al., 2025). We freeze the encoder and AnyVIA and optimize all detector components. Tab. 4 reports mAP for six encoders on all three datasets. AnyVIA improves mAP for every encoder in the table. CLIP ViT-B shows the largest gains on KUSTDrone ($+12.72$) and UAV-CB ($+16.91$). On ATR-UMOD, DINOv2-B reaches the highest mAP after adaptation among the six encoders, increasing from 10.80 to 14.96.

Table 4: **Object detection mAP (%) on KUSTDrone, UAV-CB, and ATR-UMOD.**

Dataset	Setting	DINOv2-B	DINOv3-B	CLIP ViT-B	SigLIP2-B	CLIP RN50	Swin-B
KUSTDrone	Base	45.40	45.51	26.42	43.15	34.05	28.93
	+AnyVIA	50.04 (+4.64)	48.29 (+2.78)	39.14 (+12.72)	46.64 (+3.49)	36.17 (+2.12)	33.21 (+4.28)
UAV-CB	Base	27.38	26.26	7.26	19.24	15.63	9.01
	+AnyVIA	32.55 (+5.17)	31.36 (+5.10)	24.17 (+16.91)	29.85 (+10.61)	16.50 (+0.87)	12.54 (+3.53)
ATR-UMOD	Base	10.80	8.81	0.93	9.90	1.95	1.45
	+AnyVIA	14.96 (+4.16)	12.53 (+3.72)	4.25 (+3.32)	12.32 (+2.42)	3.77 (+1.82)	2.70 (+1.25)

4.2.5 CROWD COUNTING

On DroneRGBT, the task head is a 1×1 density predictor for plain encoders. For hierarchical encoders, it aligns the last three stages and fuses them through two depthwise-separable blocks. Tab. 5 shows that AnyVIA reduces MAE and RMSE for all three encoders. SigLIP2-B has the largest reductions (6.20 MAE and 8.36 RMSE), while CLIP RN50 yields the lowest adapted errors (13.40 MAE and 15.89 RMSE).

Table 5: **Crowd counting on DroneRGBT.**

Encoder	Setting	MAE↓	RMSE↓
DINOv2-B	Base	26.41	37.09
	+AnyVIA	23.02 (-3.39)	30.74 (-6.35)
SigLIP2-B	Base	24.32	29.87
	+AnyVIA	18.12 (-6.20)	21.51 (-8.36)
CLIP RN50	Base	16.78	19.03
	+AnyVIA	13.40 (-3.38)	15.89 (-3.14)

4.3 COMPARISON WITH GENERIC FEATURE UPSAMPLING

We compare AnyVIA with the generic feature upsampler NAF (Chambon et al., 2026) on CART. Both methods use the same five frozen encoders and downstream protocol. NAF performs spatial upsampling, whereas AnyVIA combines upsampling with VI-oriented feature transformation. As shown in Fig. 3, AnyVIA improves every evaluated encoder by 0.32–2.26 mIoU points, with an average gain of 1.26 points. The gains are 0.33, 1.75, and 0.32 points for DINOv2-B, SigLIP2-B, and CLIP ViT-B, respectively. AnyVIA further improves the VI-pretrained M-SpecGene by 1.66 points and the convolutional CLIP RN50 by 2.26 points. These consistent improvements across transformer, convolutional, and VI-pretrained hosts demonstrate the value of VI-oriented feature transformation beyond spatial upsampling alone.

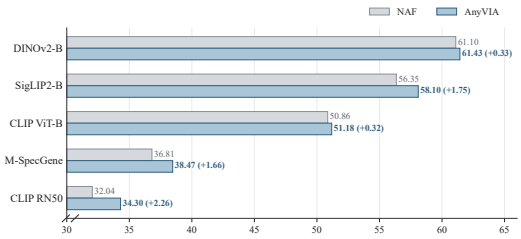


Figure 3: **Comparison with NAF on CART.**

4.4 COMPARISON WITH PER-TASK VI ADAPTATION

For frozen DINOv2-B, UniRGB-IR (Yuan et al., 2025) trains 8.90M adapter parameters per task. Tab. 6 shows that AnyVIA reuses a frozen 0.299M adapter and achieves higher performance on all three benchmarks.

Table 6: **Comparison with the encoder-specific adapter.**

Method	Trainable adapter (M)	CART mIoU↑	UAV RGB-T 2400 S_α ↑	UAV-CB mAP↑
UniRGB-IR	8.90	58.81	75.08	20.92
AnyVIA	0	**61.43**	**82.82**	**32.55**

4.5 TRANSFER TO DOMAIN-SPECIFIC FOUNDATION MODELS

We further evaluate whether AnyVIA transfers to encoders already pretrained on spectral imagery. M-SpecGene (Zhou et al., 2025) is pretrained on VI data, and InfMAE (Liu et al., 2024) is pretrained on infrared imagery. UNIP (Zhang et al., 2025a) targets infrared perception through pretraining on mixed RGB–infrared data. Tab. 7 summarizes four benchmarks spanning three downstream tasks. AnyVIA improves all three encoders on every benchmark. The largest gain for each encoder occurs on ATR-UMOD, reaching 3.80 mAP points for M-SpecGene, 3.83 for InfMAE, and 2.99 for UNIP.

Table 7: **Transfer to VI- and infrared-pretrained foundation models.**

Benchmark	Metric	M-SpecGene (VI)		InfMAE (IR)		UNIP (IR)	
		Base	+AnyVIA	Base	+AnyVIA	Base	+AnyVIA
Kust4K	mIoU↑	39.97	**40.95** (+0.98)	13.03	**13.46** (+0.43)	41.85	**43.25** (+1.40)
UAV RGB-T 2400	S_α ↑	74.22	**75.65** (+1.43)	60.16	**61.30** (+1.14)	73.99	**76.51** (+2.52)
KUSTDrone	mAP↑	46.95	**48.66** (+1.71)	37.92	**41.30** (+3.38)	46.73	**48.00** (+1.27)
ATR-UMOD	mAP↑	8.70	**12.50** (+3.80)	3.18	**7.01** (+3.83)	9.18	**12.17** (+2.99)

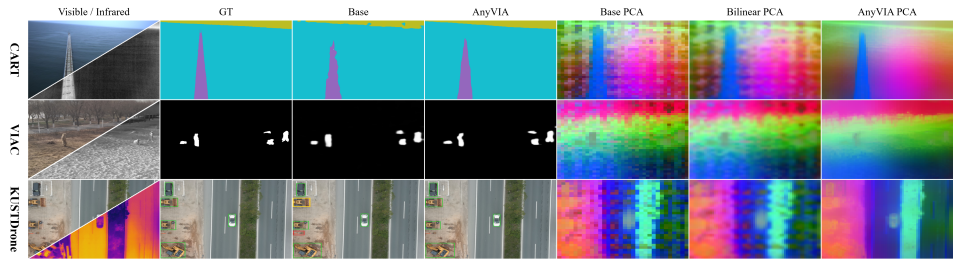


Figure 4: **Qualitative comparison.** Columns show the diagonally split visible and infrared pair, GT, the Base and AnyVIA predictions, followed by shared-basis PCA projections of the Base, bilinearly upsampled, and AnyVIA head inputs. KUSTDrone boxes mark GT and matched detections (green), false positives (red), and misses (yellow).

4.6 Feature visualization

Fig. 4 compares downstream predictions and head-input features from frozen DINOv2-B. For each example, the three PCA projections share the same basis and color range. Bilinear interpolation smooths the coarse feature grid but does not recover boundary-aligned structure. AnyVIA produces more coherent region interiors and sharper transitions, yielding predictions that better follow the CART pier and VIAC target silhouettes. On KUSTDrone, AnyVIA detects all annotated objects in the example, whereas Base adds one false positive and misses one object.

4.7 Ablation studies

Tab. 8 ablates AnyVIA on DINOv2-B under the protocol of Sec. 4.1. *Base* feeds the frozen encoder features to the task head. Every variant stays below Full AnyVIA.

Table 8: **Ablation study.**

Setting	CART mIoU↑	KUSTDrone mAP↑
Base	57.22	45.40
Bilinear 2×	57.48	45.65
w/o pretraining	54.68	37.31
w/o $\mathcal{L}_{\text{mask}}$	60.68	49.99
w/o $\mathcal{L}_{\text{crop}}$	59.32	46.70
Separate supervision	60.93	49.23
w/o g_n	61.19	49.89
Full AnyVIA	**61.43**	**50.04**

Pretraining objectives. *w/o pretraining* keeps both RGEs at their random initialization and falls below Base. *w/o* $\mathcal{L}_{\text{crop}}$ pretrains with shared-prompt mask supervision alone and loses 2.11 and 3.34 points. *w/o* $\mathcal{L}_{\text{mask}}$ pretrains with directional consistency alone and loses 0.75 and 0.05 points. *Separate supervision* scores each decoded stream independently and loses 0.50 and 0.81 points. Removing $\mathcal{L}_{\text{crop}}$ produces the larger decrease on both benchmarks. Feature-direction measurements are reported in App. C.4.

Channel folding. *w/o* g_n feeds F_m to attention at its native width and reaches 61.19 and 49.89. Folding fixes the attention width at D, allowing the learned weights to be shared across host widths.

Learned guidance versus interpolation. *Bilinear* 2× replaces AnyVIA with parameter-free interpolation at the same output resolution. It gains 0.26 and 0.25 points over Base, while AnyVIA exceeds it by 3.95 and 4.39 points (App. C.3). A matched $f = 1$ control is reported in App. C.5.

Input modalities. AnyVIA improves all three input conditions for DINOv2-B, M-SpecGene, and CLIP RN50 on CART. Across the four evaluated benchmarks, it improves all 12 paired-input settings and 17 of 24 single-modality settings (Tab. 12).

5 Conclusion

AnyVIA is an encoder-agnostic VI adapter pretrained once on UAVRGBT235K and reused in frozen form across vision encoders and aerial tasks. Raw-image guidance decouples its learned parameters from the host feature space. Shared-prompt supervision coordinates the two modalities, while directional consistency aligns refined and host feature directions. Across eight held-out UAV benchmarks and five tasks, AnyVIA improves diverse pretrained encoders.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript writing and language polishing, including improvements to grammar, wording, sentence structure, clarity, and readability. All AI-assisted text was critically reviewed, revised, and verified by the authors against the underlying methods, experimental records, and cited sources. The authors take full responsibility for the accuracy, originality, claims, and final content of the manuscript.

REPRODUCIBILITY STATEMENT

Sections 3.2–3.4 describe the construction of UAVRGBT235K, the AnyVIA architecture, and the pretraining objectives. Section 4.1 specifies the downstream benchmarks, metrics, and matched evaluation protocol. Appendix A reports the per-source dataset composition and mask-generation pipeline, while Appendix B gives the module configuration, optimization settings, and feature–crop alignment. Appendix C presents the additional transfer, scaling, control, modality, ablation, and efficiency analyses.

REFERENCES

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026.

Loïck Chambon, Paul Couairon, Eloi Zablocki, Alexandre Boulch, Nicolas Thome, and Matthieu Cord. Naf: Zero-shot feature upsampling via neighborhood attention filtering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26604–26613, 2026.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27958–27967. IEEE, 2025.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258, 2017.

Paul Couairon, Loick Chambon, Louis Serrano, Jean-Emmanuel Haugeard, Matthieu Cord, and Nicolas Thome. Jafar: Jack up any feature at any resolution. *Advances in Neural Information Processing Systems*, 38:99265–99292, 2025.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024.

Wenhui Diao, Haichen Yu, Kaiyue Kang, Tong Ling, Di Liu, Yingchao Feng, Hanbo Bi, Libo Ren, Xuexue Li, Yongqiang Mao, et al. Ringmo-aerial: An aerial remote sensing foundation model with affine transformation contrastive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yuchuan Ding, Linfei Li, Lin Zhang, and Ying Shen. Raysup: Ultra-light universal feature upsampling via geometry-aware ray representation. In *European Conference on Computer Vision*, 2026.

Hangtao Feng, Lu Zhang, Siqi Zhang, Dong Wang, Xu Yang, and Zhiyong Liu. Rtdod: A large-scale rgb-thermal domain-incremental object detection dataset for uavs. *Image and Vision Computing*, 140:104856, 2023.

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27672–27683, 2024.

Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6185–6194. IEEE, 2023.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pp. 289–305. Springer, 2024.

Shenghui Huang, Menghao Hu, Longkun Zou, Hongyu Chi, Zekai Li, Feng Gao, Fan Yang, Qingyao Wu, and Ke Chen. Uav-cb: A complex-background rgb-t dataset and local frequency bridge network for uav detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40468–40478, 2026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Connor Lee, Matthew Anderson, Nikhil Ranganathan, Xingxing Zuo, Kevin Do, Georgia Gkioxari, and Soon-Jo Chung. Caltech aerial rgb-thermal dataset in the wild. In *European Conference on Computer Vision*, pp. 236–256. Springer, 2024.

Bingyu Li, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Stitchfusion: Weaving any visual modalities to enhance multimodal semantic segmentation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1308–1317, 2025. doi: 10.1145/3746027.3755110.

Cheng Liu, Zheng Wang, Xinyu Yan, Meijun Sun, and Qinghua Hu. Visible-infrared camouflaged object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2):1716–1728, 2026. doi: 10.1109/TCSVT.2025.3608933.

Fangcen Liu, Chenqiang Gao, Yaming Zhang, Junjie Guo, Jinghao Wang, and Deyu Meng. Infmae: A foundation model in the infrared modality. In *European Conference on Computer Vision*, pp. 420–437. Springer, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Maxime Oquab, Timothee Darcet, Theo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Junlin Ouyang, Qingwang Wang, Ying Shang, Pengcheng Jin, Hangwei Zhong, Liman Zhou, and Tao Shen. An rgb-tir dataset from uav platform for robust urban traffic scenes semantic segmentation. *Scientific Data*, 12(1):1701, 2025.

Tao Peng, Qing Li, and Pengfei Zhu. Rgb-t crowd counting from drone: A benchmark and mmccn network. In *Asian Conference on Computer Vision*, pp. 497–513. Springer, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*, 2025.

Sebastien Razakarivony and Frederic Jurie. Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34:187–203, 2016.

Isaac Robinson, Peter Robicheaux, Matvei Popov, Deva Ramanan, and Neehar Peri. Rf-detr: Neural architecture search for real-time detection transformers. *arXiv preprint arXiv:2511.09554*, 2025.

Oriane Simeoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michael Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jiamin Song, Nan Zhang, Zhenhao Wang, and Tian Tian. Escvehicle: A drone-based visible–infrared vehicle benchmark with extensive scene coverage. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–16, 2026. doi: 10.1109/TGRS.2026.3655959.

Kechen Song, Hongwei Wen, Xiaotong Xue, Liming Huang, Yingying Ji, and Yunhui Yan. Modality registration and object search framework for uav-based unregistered rgb-t image salient object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

Simon Speth, Artur Gonçalves, Bastien Rigault, Satoshi Suzuki, Mondher Bouazizi, Yutaka Matsuo, and Helmut Prendinger. Deep learning with rgb and thermal images onboard a drone for monitoring operations. *Journal of Field Robotics*, 39(6):840–868, 2022.

Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, et al. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–22, 2022a.

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

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Xiaohua Zhai, Basil Mustafa, Lucas Beyer, Andreas Steiner, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Shashanka Venkataramanan, Valentinos Pariza, Mohammadreza Salehi, Lukas Knobel, Elias Ramzi, Spyros Gidaris, Andrei Bursuc, and Yuki M. Asano. Franca: Nested matryoshka clustering for scalable visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10533–10544, 2026.

Qingwang Wang, Yuxuan Sun, Tao Shen, Mohammad Al-Antary, Hisham Alasmary, and Muhammad Waqas. Mgsanet: A multiscale graph spatial alignment network for weakly aligned rgb-thermal object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–18, 2026. doi: 10.1109/TGRS.2025.3647051.

Thomas Wimmer, Prune Truong, Marie-Julie Rakotosaona, Michael Oechsle, Federico Tombari, Bernt Schiele, and Jan Eric Lenssen. Anyup: Universal feature upsampling. In *International Conference on Learning Representations*, 2026.

Yun Xiao, Dan Cao, Chenglong Li, Bo Jiang, and Jin Tang. A benchmark dataset for high-altitude uav multi-modal tracking. *Journal of Image and Graphics*, 30(02):0361–0374, 2025. doi: 10.11834/jig.240040.

Xinyi Ying, Chao Xiao, Wei An, Ruojing Li, Xu He, Boyang Li, Xu Cao, Zhaoxu Li, Yingqian Wang, Mingyuan Hu, et al. Visible-thermal tiny object detection: A benchmark dataset and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):6088–6096, 2025.

Maoxun Yuan and Xingxing Wei. C^2former: Calibrated and complementary transformer for rgb-infrared object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

Maoxun Yuan, Bo Cui, Tianyi Zhao, Jiayi Wang, Shan Fu, Xue Yang, and Xingxing Wei. Unirgb-ir: A unified framework for visible-infrared semantic tasks via adapter tuning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2409–2418, 2025.

Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 24(12):14679–14694, 2023a. doi: 10.1109/TITS.2023.3300537.

Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8876–8885. IEEE, 2022.

Tao Zhang, Jinyong Wen, Zhen Chen, Kun Ding, Shiming Xiang, and Chunhong Pan. Unip: Rethinking pre-trained attention patterns for infrared semantic segmentation. In *International Conference on Learning Representations*, 2025a.

Xiangqing Zhang, Yan Feng, Nan Wang, Guohua Lu, and Shaohui Mei. Aerial person detection for search and rescue: Survey and benchmarks. *Journal of Remote Sensing*, 5:0474, 2025b.

Yan Zhang, Chang Xu, Wen Yang, Guangjun He, Huai Yu, Lei Yu, and Gui-Song Xia. Drone-based rgbt tiny person detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 204:61–76, 2023b.

Yi Zhang, Wang Zeng, Sheng Jin, Chen Qian, Ping Luo, and Wentao Liu. When pedestrian detection meets multi-modal learning: Generalist model and benchmark dataset. In *European Conference on Computer Vision*, 2024.

Kailai Zhou, Fuqiang Yang, Shixian Wang, Bihan Wen, Chongde Zi, Linsen Chen, Qiu Shen, and Xun Cao. M-specgene: Generalized foundation model for rgbt multispectral vision. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7861–7872. IEEE, 2025.

A UAVRGBT235K IN DETAIL

Tab. 9 lists the ten sources and their contribution, and Fig. 5 shows the annotation pipeline summarized in Sec. 3.2. Eight detection sources contribute 105,175 pairs and two tracking sources the remaining 12,493. The prompted branch converts each source box or oriented box into an instance mask. The automatic branch supplies additional unprompted object masks. Their merged outputs yield 60.8 masks per pair.

Table 9: **UAVRGBT235K composition.** Ten aerial visible–infrared sources are standardized into paired, modality-wise panoptic masks. *None* of the eight datasets used for downstream evaluation is included. All downstream results in this paper are out-of-domain with respect to adapter pretraining.

Source	Task	Pairs	Masks	Masks/pair
RGBT-Tiny (Ying et al., 2025)	detection	34,015	1,414,137	41.6
DroneVehicle (Sun et al., 2022b)	detection	28,437	1,692,321	59.5
RTDOD (Feng et al., 2023)	detection	16,176	1,119,656	69.2
ESCVehicle (Song et al., 2026)	detection	10,727	733,995	68.4
VTUAV (Zhang et al., 2022)	tracking	7,057	476,380	67.5
RGBTDronePerson (Zhang et al., 2023b)	detection	6,125	534,130	87.2
HiAI (Xiao et al., 2025)	tracking	5,436	859,157	158.1
VTSaR (Zhang et al., 2025b)	detection	5,324	170,277	32.0
NII-CU (Speth et al., 2022)	detection	3,126	80,482	25.7
VEDAI (Razakarivony & Jurie, 2016)	detection	1,245	73,981	59.4
UAVRGBT235K		**117,668**	**7,154,516**	**60.8**

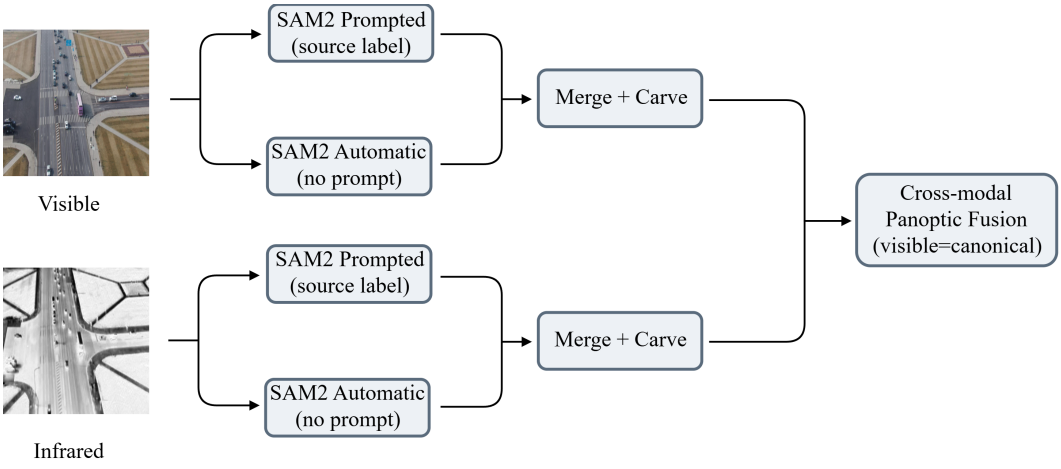


Figure 5: **UAVRGBT235K annotation pipeline.** A frozen SAM2 (Ravi et al., 2025) processes the visible and infrared images independently through two complementary branches. The prompted branch converts source labels into instance masks, whereas the automatic branch recovers objects absent from the source annotations. Within each modality, the two mask sets are merged and carved into a panoptic layer. The visible and infrared layers are then fused in the visible coordinate frame to produce cross-modal object-level supervision for AnyVIA pretraining.

B ANYVIA IMPLEMENTATION DETAILS

B.1 RAW GUIDANCE ENCODER

Each RGE uses a 3×3 stem convolution to project the three-channel guidance image to $D = 128$ channels. The detail branch applies GroupNorm, SiLU, a depthwise 3×3 convolution, SiLU, and a 1×1 pointwise projection to produce $D/2$ channels. The context branch first applies a 1×1 GSC

projection to $D/2$ channels, followed by three 3×3 GSC blocks. The initial projection and the three cascade outputs are concatenated and reduced from $2D$ to $D/2$ channels by a 1×1 CGS block. Concatenating the detail and context outputs restores the channel width to D. Adaptive average pooling maps the result to the guidance resolution, after which 2D rotary positional encoding (Heo et al., 2024) encodes spatial position.

B.2 Training settings

We pretrain AnyVIA for 10 epochs using AdamW with an initial learning rate of 2×10^{-4}, cosine decay to 2×10^{-5}, and weight decay of 10^{-5}. Training uses 8 GPUs with 8 image pairs per GPU, giving a global batch size of 64. The host encoder receives 512×512 images. We sample up to three instance masks per image pair and derive a shared box prompt for the two modality branches from each mask. The focal, Dice, IoU, and object-score terms in the SAM2 objective are weighted by 20, 1, 1, and 1, respectively. Mask logits are summed before mask supervision, whereas the auxiliary IoU and object-score predictions are averaged across modalities. Pretraining updates the two RGEs. The Hiera-B+ trunk, FPN neck, prompt encoder, and mask decoder remain frozen.

B.3 Feature and crop alignment

For each modality, AnyVIA refines the stride-16 Hiera-B+ trunk feature map ($C = 448$) from 32×32 to 64×64. Directional consistency is evaluated at the refined resolution. For mask supervision, the refined features are average-pooled to 32×32 before entering the frozen FPN neck and mask decoder. Preprocessing retains I_m^{hr} at 1024×1024 and forms the 512×512 encoder input I_m. A 256×256 reference crop from I_m^{hr} is encoded at its native resolution to produce a 16×16 feature map. The corresponding window in the refined 64×64 map also has a resolution of 16×16, covering the same image region at the same spatial sampling density. Crop coordinates are selected from the aggregate foreground mask, constrained to the image boundaries, and shared across modalities.

C Additional Analyses

C.1 Transfer across encoder scales

We evaluate AnyVIA with the small, base, and large variants of DINOv2. As shown in Tab. 10, the same frozen adapter improves all three encoder scales on five benchmarks, covering semantic segmentation, salient object detection, and horizontal and oriented object detection. The gains are positive in all 15 settings and range from 0.66 to 4.64 points.

Table 10: **Transfer across DINOv2 encoder scales.** Results are reported on the five benchmarks for which AnyVIA improves all three encoder scales. Each parenthesized value denotes the improvement over the corresponding Base result.

Encoder	Setting	CART mIoU↑	Kust4K mIoU↑	UAV RGB-T 2400 S_α ↑	KUSTDrone mAP↑	ATR-UMOD mAP↑
DINOv2-S	Base	54.16	51.47	78.89	44.40	8.72
	+AnyVIA	**55.77** (+1.61)	**53.52** (+2.05)	**81.13** (+2.24)	**48.27** (+3.87)	**12.88** (+4.16)
DINOv2-B	Base	57.22	55.34	80.22	45.40	10.80
	+AnyVIA	**61.43** (+4.21)	**58.16** (+2.82)	**82.82** (+2.60)	**50.04** (+4.64)	**14.96** (+4.16)
DINOv2-L	Base	61.26	57.30	80.60	51.37	14.14
	+AnyVIA	**63.37** (+2.11)	**60.08** (+2.78)	**83.11** (+2.51)	**52.03** (+0.66)	**16.71** (+2.57)

C.2 Effect of pretraining data scale

We pretrain AnyVIA on 25%, 50%, and 100% of UAVRGBT235K and evaluate each frozen adapter with DINOv2-B under the downstream protocols used in the main experiments. Performance increases with the amount of pretraining data on both CART and KUSTDrone (Fig. 6). At 25% of the data, AnyVIA retains 89% and 60% of its full-data gains on CART and KUSTDrone, respec-

tively. These ratios increase to 95% and 81% with 50% of the data. Performance rises monotonically through the full-data setting on both benchmarks.

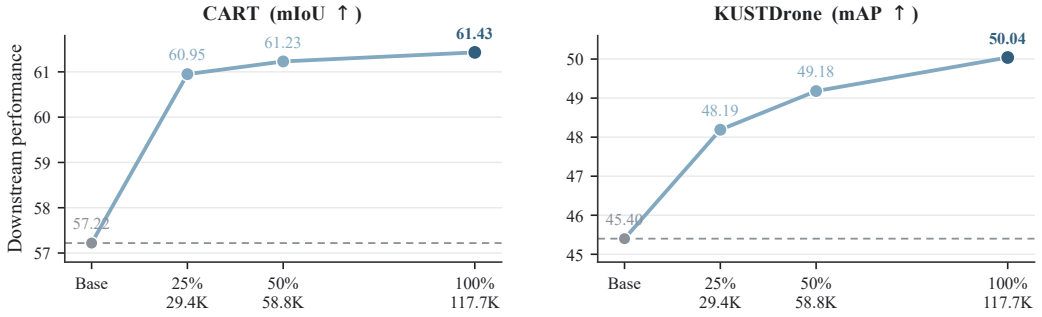


Figure 6: **Effect of pretraining data scale.** Downstream performance with DINOv2-B after pretraining AnyVIA on 29,417–117,668 image pairs.

Table 11: **Comparison with bilinear 2× upsampling.** Base uses the original feature resolution, while bilinear upsampling and AnyVIA produce features at the same 2× resolution. SOD is evaluated by S_α and detection by mAP. Δ denotes the improvement of AnyVIA over bilinear upsampling.

Benchmark	Encoder	Base 1×	Bilinear 2×	AnyVIA 2×	Δ vs. bilinear
Salient object detection					
UAV RGB-T 2400	DINOv2-B	80.22	80.64	**83.10**	+2.46
	M-SpecGene	74.22	75.02	**75.99**	+0.97
Horizontal object detection					
KUSTDrone	DINOv2-B	45.49	45.75	**49.84**	+4.09
	CLIP ViT-B	26.42	26.47	**39.14**	+12.67
	Swin-B	28.93	29.52	**33.21**	+3.69
	CLIP RN50	34.05	34.97	**36.17**	+1.20
UAV-CB	DINOv2-B	27.38	28.14	**32.55**	+4.41
	CLIP ViT-B	7.26	8.05	**24.17**	+16.12
	Swin-B	9.01	9.60	**12.54**	+2.94
	CLIP RN50	15.63	16.19	**16.50**	+0.31
Oriented object detection					
ATR-UMOD	DINOv2-B	10.80	13.14	**14.96**	+1.82
	CLIP ViT-B	0.93	0.85	**4.25**	+3.40
	Swin-B	1.45	1.62	**2.70**	+1.08
	CLIP RN50	1.95	2.53	**3.77**	+1.24

C.3 Bilinear interpolation control

We compare AnyVIA with parameter-free bilinear 2× upsampling in matched control runs. Base, bilinear upsampling, and AnyVIA use the same frozen encoder, task-head architecture, and training protocol. As shown in Tab. 11, AnyVIA outperforms bilinear upsampling in all 14 settings, with gains ranging from 0.31 to 16.12 points. The largest margin occurs with CLIP ViT-B on UAV-CB, and the smallest with CLIP RN50 on UAV-CB.

C.4 Feature-direction consistency

We evaluate feature-direction consistency on the CART validation set for three unseen encoders using the distance in Eq. 7. Fig. 7 compares bilinear 2× upsampling with AnyVIA at factors $f = 1$ and $f = 2$. The full $f = 2$ model produces the lowest distance for both modalities with every encoder. Averaged over modalities, it reduces the distance from bilinear interpolation by 0.0077, 0.0397, and 0.0669 for DINOv2-B, CLIP ViT-B, and SigLIP2-B, respectively. Across the paired

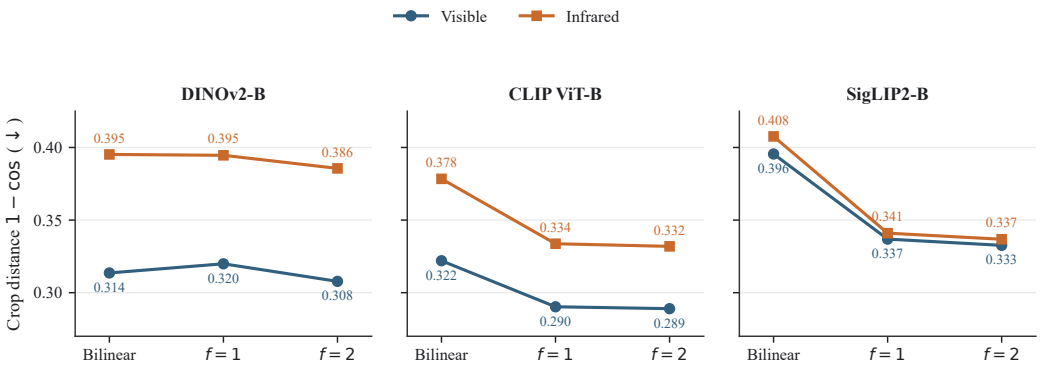


Figure 7: **Feature-direction consistency on the CART validation set.** Curves show the mean $1 -$ cos distance between refined crop features and frozen host features extracted from the corresponding higher-resolution crops. Results are reported separately for visible and infrared inputs. Lower values indicate stronger consistency.

comparisons, $f = 2$ yields a lower distance than bilinear interpolation in 93.3% of cases and than $f = 1$ in 92.3%.

C.5 Feature transformation and spatial upsampling

The $f = 1$ variant preserves the original feature resolution and performs feature transformation without spatial upsampling. It improves Base by 3.98 mIoU on CART and 0.13 S_α on VIAC, while reducing MAE by 3.71 on DroneRGBT (Fig. 8). The $f = 1$ and $f = 2$ results differ by 0.23 mIoU, 0.04 S_α, and 0.32 MAE on the three benchmarks, respectively.

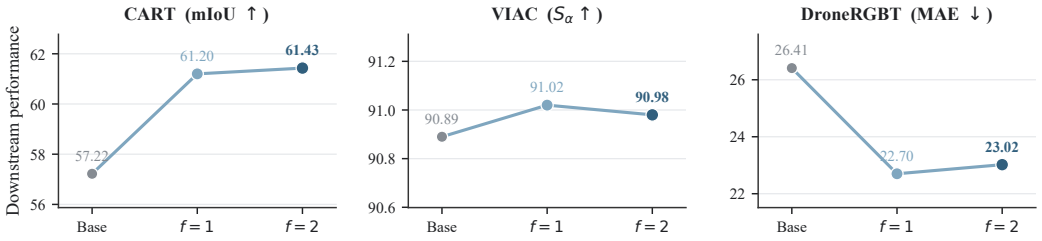


Figure 8: **Feature transformation and spatial upsampling.** Absolute performance is reported for Base, AnyVIA at the original feature resolution ($f = 1$), and AnyVIA with $2\times$ spatial upsampling ($f = 2$). Lower MAE is better.

C.6 Input modality analysis

We evaluate Base and AnyVIA with visible-only, infrared-only, and paired inputs for three encoders across four benchmarks. For UAV-CB, each single-modality setting uses the annotations associated with that modality. As shown in Tab. 12, AnyVIA improves all 12 paired-input settings and 17 of the 24 single-modality settings. The paired-input gains range from 0.87 to 5.17 points across the evaluated encoders and benchmarks.

Table 12: **Input modality analysis across four benchmarks.** Base and AnyVIA are evaluated with visible-only, infrared-only, and paired inputs using DINOv2-B, M-SpecGene, and CLIP RN50. Each cell reports Base/AnyVIA.

Encoder	Input	CART mIoU	Kust4K mIoU	KUSTDrone mAP	UAV-CB mAP
DINOv2-B	Visible	55.77 / **59.24**	**48.10**/48.03	43.90 / **47.77**	26.59 / **27.47**
	Infrared	52.24 / **54.15**	43.17 / **43.23**	35.07 / **37.06**	**33.20**/33.04
	Paired	57.22 / **61.43**	55.34 / **58.16**	45.40 / **50.04**	27.38 / **32.55**
M-SpecGene	Visible	35.46 / **35.62**	**34.65**/33.54	**46.45**/46.31	**24.03**/21.83
	Infrared	26.58 / **26.90**	**27.79**/26.60	36.52 / **36.88**	36.49 / **38.15**
	Paired	37.34 / **38.47**	39.97 / **40.95**	46.95 / **48.66**	28.10 / **31.13**
CLIP RN50	Visible	32.94 / **35.04**	34.75 / **35.86**	31.95 / **33.10**	**12.67**/10.03
	Infrared	23.30 / **25.44**	26.92 / **27.18**	25.53 / **27.67**	17.09 / **18.22**
	Paired	31.39 / **34.30**	35.88 / **38.74**	34.05 / **36.17**	15.63 / **16.50**

C.7 Guidance source

AnyVIA derives Q and K from the input image. A matched variant derives them from the folded host features. This variant reaches 60.22 on CART and 42.06 on KUSTDrone with 0.59M adapter parameters, compared with 61.43, 50.04, and 0.299M for AnyVIA (Tab. 13).

Table 13: **Guidance source.** Performance and adapter size when Q and K are derived from the input image or folded host features.

Q and K derived from	Adapter params	CART	KUSTDrone
Input image (AnyVIA)	0.299M	**61.43**	**50.04**
Folded host features	0.59M	60.22	42.06

C.8 Computational efficiency

We measure the adapter-only efficiency of AnyVIA with DINOv2-S/B/L at output scales $f \in \{1, 2\}$ (Tab. 14). Measurements use paired visible and infrared inputs, batch size 1, and FP32 on an RTX 4090. The host feature maps are 32×32, the raw images are 448×448, and the two output scales correspond to 32×32 and 64×64. Latency is the median of 200 CUDA-event measurements after 50 warm-up iterations.

AnyVIA retains 0.299M learned parameters as the host width increases from 384 to 1024. At $f = 1$, computation ranges from 29.4 to 78.3 GFLOPs, latency from 4.31 to 7.25 ms, and peak memory from 174 to 454 MiB. At $f = 2$, the corresponding ranges are 117.4–313.1 GFLOPs, 14.98–38.84 ms, and 648–1718 MiB.

Table 14: **Computational efficiency of AnyVIA across encoder scales and output resolutions.** C, n, and f denote host width, channel-group count, and output scale, respectively.

Encoder	C	n	Output	Adapter params (M)	GFLOPs	Latency (ms)	Peak memory (MiB)
DINOv2-S	384	3	1×	0.299	29.4	4.31	174
			2×		117.4	14.98	648
DINOv2-B	768	6	1×		58.7	5.22	342
			2×		234.8	28.87	1290
DINOv2-L	1024	8	1×		78.3	7.25	454
			2×		313.1	38.84	1718