

AVID: A Weakly Aligned RGB–Infrared Dataset and Perturbation Benchmark for Tiny Aerial Targets

Dharini Raghavan¹ Vishnu Rao²

¹Georgia Institute of Technology ²ARTPARK, Indian Institute of Science

draghavan7@gatech.edu

 Code

 Hugging Face

Abstract

RGB and infrared sensing are naturally complementary for aerial surveillance, but their value is harder to assess when targets are very small, the two views are only weakly aligned, and each modality responds differently to image degradation. We introduce AVID, a paired RGB–infrared dataset and benchmark for tiny aerial targets that makes these factors explicit. The dataset contains 32,040 paired frames from 115 short surveillance videos with 35,952 annotated bird and drone instances. The target scale is unusually small: 68.1% of all instances have a geometric-mean side length between 8 and 16 pixels in a 320×256 frame, and 99.0% fall within the COCO “small” category.

We also study the geometric agreement between the paired sensors. On a sample of 300 RGB–infrared frame pairs, the median estimated shift is 1 pixel, while 32.3% of pairs exceed 5 pixels. Because the same pixel displacement can have very different consequences for small and large objects, we introduce the object-relative alignment ratio $\rho_{\text{align}} = d_{\text{align}}/s_{\text{target}}$ and the corresponding alignment burden curve $\hat{A}(\tau) = \Pr(\rho_{\text{align}} > \tau)$. In the measured sample, 33.0% of pairs exceed one quarter of a target side, 29.7% exceed one half, and 16.7% exceed one full target side, giving a unit-scale burden of 0.304.

To study sensor degradation, we introduce AVID-C, a 7×5 perturbation benchmark covering Gaussian, speckle, and salt-and-pepper noise, Gaussian and motion blur, uneven illumination, and camera instability. The perturbations can be applied separately to RGB and infrared imagery, allowing modality-specific sensitivity to be examined directly. We further characterize how target scale, corruption, and cross-modal shift change the overlap between bird and drone distributions using Bhattacharyya and Chernoff error bounds, a kernel-density cross-check, a fusion-gap measure, and mutual-information sensitivity.

Together, AVID and AVID-C provide a benchmark in which target scale, cross-modal alignment, and sensor degradation are measured explicitly rather than being absorbed into a single aggregate task score. We hope this makes the dataset useful not only for comparing methods, but also for understanding when and why multimodal sensing helps in the tiny-target regime.

1 Introduction

RGB and thermal cameras provide complementary views of the same scene. Visible imagery preserves color, texture, and fine spatial structure, while thermal imagery captures differences in emitted radiation and can remain informative when visible contrast is weak. This complementarity is particularly useful for long-range aerial surveillance, where drones may occupy only a small number of pixels and can be difficult to distinguish from birds or background structure. The IEEE ICIP 2025 Video and Image Processing Cup considered this setting through a challenge on infrared–visual drone surveillance [1]. The data collected for that effort form the basis of the dataset studied here.

A paired RGB–infrared dataset, however, is not fully characterized by the number of image pairs it contains. For distant targets, the scale of the object, the geometric agreement between the two sensors, and the quality of each sensor observation can all influence how much useful information is available. These factors become especially important when the target is very small. A displacement of five pixels is minor for a large object, but it can correspond to nearly half the width of a small drone. Similarly, blur or noise that appears mild at the scale of the full image can remove a large fraction of the visual structure within a target that is only 8 or 12 pixels across. The effect can also differ between RGB and thermal imagery because the two sensors respond to different physical signals.

We introduce AVID, a paired RGB–infrared dataset and benchmark for tiny aerial targets that makes these properties visible rather than treating them as hidden details of the data. The paired release contains 32,040 RGB–IR frame pairs from 115 short surveillance videos and 35,952 annotated bird and drone instances. The dataset occupies an unusually small-object regime: 24,489 instances, or 68.1% of the corpus, have a geometric-mean box side between 8 and 16 pixels in a 320×256 frame, and 99.0% of all annotated instances fall within the COCO “small” category. Birds account for 15,471 instances and drones for 20,481.

The small target size also changes how cross-modal alignment should be described. A displacement measured only in pixels does not have the same meaning across different target sizes. We therefore measure alignment relative to the object scale in each frame. For a paired frame i , let d_i denote the estimated global RGB–infrared displacement and let s_i denote the median geometric-mean side length of the annotated targets in that frame. We define the object-relative alignment ratio as

$$\rho_i = \frac{d_i}{s_i}. \quad (1)$$

The ratio expresses the sensor displacement in units of target size. For example, $\rho_i = 0.5$ means that the estimated displacement is equal to one half of the median target side in that frame, while $\rho_i > 1$ means that it is larger than one target side. This makes the same alignment measurement easier to interpret across scenes with different object scales.

A single mean or median does not describe how often large relative shifts occur. We therefore define the alignment burden curve

$$\hat{A}(\tau) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}[\rho_i > \tau], \quad (2)$$

which gives the fraction of frame pairs whose estimated displacement exceeds a fraction τ of the target side. We also define a truncated summary

$$\hat{B}_T = \frac{1}{T} \int_0^T \hat{A}(\tau) d\tau = \frac{1}{nT} \sum_{i=1}^n \min(\rho_i, T). \quad (3)$$

We use $T = 1$ when summarizing displacement up to one target side. On a sample of 300 paired frames, the median estimated RGB–infrared shift is 1 pixel, while 32.3% of the pairs exceed 5 pixels. When the displacement is expressed relative to target size, 33.0% of the pairs exceed one quarter of a target side, 29.7% exceed one half, and 16.7% exceed one full target side. The corresponding unit-scale burden is $\hat{B}_1 = 0.304$. The alignment search is bounded at ± 12 pixels along each image axis, so the upper end of the measured distribution is censored and should not be interpreted as a physical limit on the sensor offset.

We complement the dataset with AVID-C, a perturbation benchmark designed for the native 320×256 imagery. It contains seven perturbation families—Gaussian noise, speckle noise, salt-and-pepper noise, Gaussian blur, uneven illumination, motion blur, and camera instability—with five severity levels for each family. The perturbations can be applied to RGB, infrared, or both modalities. This allows the benchmark to distinguish between degradation of the visible stream, degradation of the thermal stream, and changes that affect both observations.

Beyond describing the perturbations themselves, we study how the underlying bird and drone distributions change with target scale, corruption, and cross-modal displacement. We use a fixed low-dimensional representation of annotated object patches and measure class overlap with Bhattacharyya and Chernoff error upper bounds, together with a kernel-density cross-check. We also introduce a fusion-gap diagnostic to summarize the difference between paired and single-modality class overlap, and we examine how this quantity changes as the infrared observation is shifted relative to RGB. These measurements are not intended to replace downstream task evaluation. Instead, they provide a complementary view of the dataset by showing how much class structure remains under different data conditions.

Taken together, AVID is intended to support a more informative style of multimodal evaluation. A result on this dataset can be interpreted together with the target scales present in the evaluation, the degree of RGB–infrared alignment, and the type and severity of sensor degradation. This makes it easier to understand which parts of the data distribution are challenging and where the two sensing modalities provide complementary information.

The main contributions of this work are:

1. **A paired RGB–infrared dataset for tiny aerial targets.** AVID contains 32,040 verified RGB–IR frame pairs from 115 surveillance videos with 35,952 annotated bird and drone instances. The dataset is dominated by very small targets, with 68.1% of instances falling between 8 and 16 pixels in geometric-mean side length.
2. **A scale-aware description of weak cross-modal alignment.** We introduce the object-relative alignment ratio in Eq. 1, the alignment burden curve in Eq. 2, and the truncated burden score in Eq. 3. These quantities express sensor displacement relative to the size of the target rather than only in image pixels.
3. **AVID-C, a multimodal perturbation benchmark.** We define seven perturbation families at five severity levels and allow them to be applied separately to RGB and infrared imagery. This makes modality-specific sensitivity to noise, blur, illumination changes, motion blur, and camera instability part of the benchmark.

4. **A data-centered analysis of scale, corruption, and alignment.** We characterize how bird–drone class overlap changes across target scale, perturbation severity, and controlled cross-modal shift, using Bhattacharyya and Chernoff error bounds, a kernel-density cross-check, a fusion-gap measure, and mutual-information sensitivity.

2 Related datasets and benchmark design

2.1 Visible–infrared and UAV datasets

Visible and thermal sensing has been studied extensively in pedestrian detection, autonomous driving, tracking, and scene understanding. KAIST introduced synchronized multispectral pedestrian imagery and established an early benchmark for combining visible and thermal observations [2]. LLVIP focuses on paired visible–infrared imagery under low-light conditions [3], while M3FD contains visible–infrared scenes spanning multiple environments and tasks, including image fusion and object detection [4]. FLIR ADAS similarly provides visible and thermal driving imagery for perception in road scenes [5]. These datasets demonstrate the value of complementary sensing, but they primarily study objects that occupy a larger fraction of the image than the distant aerial targets considered here.

A separate line of work has developed datasets for UAV perception. VisDrone studies objects observed from cameras mounted on drones [6], whereas AVID considers the opposite setting: the drone is the distant object being observed from a surveillance camera. Visible–thermal UAV datasets have also been introduced for related tasks. DroneRGBT studies RGB–thermal aerial crowd counting [7], and VTUAV provides paired visible–thermal video for tracking [8]. Anti-UAV focuses more directly on surveillance of unmanned aerial vehicles and includes RGB and infrared video data [9, 10].

A related line of work has studied the perception problem itself. Sethu Selvi et al. considered visual discrimination between drones and birds using deep object detectors [11] and also explored the fusion of RGB and IR modalities for drone localization [12]. Raghavan et al. studied drone localization and tracking in real-time surveillance video [13], and further extended this setting to infrared imagery under low visibility, weak contrast, uneven illumination, and image distortion [14]. These studies highlight two difficulties that are central to long-range aerial surveillance: distinguishing drones from visually similar birds and maintaining useful target information as imaging conditions deteriorate. AVID takes a complementary, data-centered view by characterizing the target scale, RGB–infrared alignment, and modality-specific degradation present in paired aerial observations.

More recent datasets have expanded the range of multimodal UAV sensing. MMFW-UAV studies fixed-wing UAVs with multiple sensors and viewpoints [15]; MM-UAV combines RGB, infrared, and event cameras for UAV tracking [16]; UAVNet-MS uses RGB and multispectral imagery for fine-grained recognition of small UAVs [17]; and UAV-CB studies RGB–thermal UAV observations under complex backgrounds and camouflage [18]. Together, these datasets cover a broad range of sensing configurations, target types, and downstream tasks.

AVID occupies a complementary part of this landscape. It focuses on bird–drone discrimination in surveillance imagery where the targets are consistently very small and the RGB and infrared views are paired but not assumed to be perfectly registered. The dataset is therefore organized around three properties that become especially important in this regime: target scale, cross-modal alignment, and sensor degradation. Rather than treating these as secondary characteristics of the data, we measure them directly and make them part of the benchmark.

Table 1 summarizes the closest related datasets and highlights the aspects that are most relevant to the setting studied here.

2.2 Corruption and data-centric benchmarks

Robustness benchmarks have shown that a model’s behavior can change substantially when the input distribution is altered in controlled ways. ImageNet-C established a widely used framework in which common image corruptions are grouped into families and evaluated across increasing severity levels [19]. Similar ideas have since been applied to aerial and multimodal settings. UAV-C studies corruption robustness for aerial tracking [20], while MapBench examines how different sensing modalities respond to environmental and sensor perturbations in multi-sensor map construction [21]. These works show that robustness is often easier to interpret when the type and strength of the degradation are stated explicitly rather than collapsed into a single average score.

AVID-C follows this general idea but is designed around the sensing regime of AVID. Its seven perturbation families are defined for the native 320×256 RGB–thermal imagery and include noise, blur, uneven illumination, motion blur, and camera instability. Importantly, the two modalities can be perturbed separately. This matters for multimodal data because degradation need not affect both sensors in the same way, and an aggregate corruption score can hide which modality has lost useful information.

Our benchmark is also related to recent work that treats the properties of a dataset as part of the evaluation itself. DataPerf studies data quality and data-centric learning as benchmark variables [22], while DataComp evaluates how dataset construction influences downstream behavior [23]. WhyShift emphasizes that distribution shift is more informative when the source of

Table 1: Representative visible–infrared and UAV datasets related to AVID. The “Pairing” column describes how the modalities are provided by each dataset and should not be interpreted as pixel-level registration unless stated by the corresponding source.

Dataset	Main target/view	Modalities	Pairing	Main characteristic
KAIST [2]	pedestrian, ground	RGB + thermal	synchronized	multispectral pedestrian benchmark
LLVIP [3]	pedestrian, ground	visible + IR	paired	low-light visible–infrared imagery
M3FD [4]	multi-class, ground	visible + IR	paired	multi-scene fusion and detection
VTUAV [8]	generic target, aerial	visible + thermal	paired video	visible–thermal tracking
Anti-UAV300 [10]	UAV, surveillance	RGB + IR	both streams released	long-range anti-UAV surveillance
MMFW-UAV [15]	fixed-wing UAV, air-to-air	visible + thermal	one-to-one	multi-sensor and multi-view UAV imagery
MM-UAV [16]	UAV tracking	RGB + IR + event	synchronized	tri-modal UAV sequences
UAVNet-MS [17]	small UAV	RGB + MSI	synchronized	fine-grained multispectral recognition
UAV-CB [18]	UAV, low-altitude	RGB + thermal	paired	complex backgrounds and camouflage
AVID (ours)	bird/drone, surveillance	RGB + thermal	weakly aligned pairs	tiny targets, measured alignment, and controlled perturbations

the shift is identified rather than represented only by a change in accuracy [24]. Mayo et al. similarly study visual dataset difficulty directly rather than relying only on the performance of a particular recognition system [25].

AVID adopts the same broader view in a multimodal setting. We characterize the data along dimensions that are especially important for tiny RGB–infrared targets: how much spatial evidence is available, how far the two sensor observations are displaced relative to the target size, and how each modality changes under controlled degradation. The goal is not to replace downstream evaluation, but to give those results a clearer context by making the condition of the evaluated data explicit.

3 Dataset construction and release scope

3.1 Source and scope

The source imagery was collected for the IEEE Signal Processing Society Video and Image Processing Cup at ICIP 2025, “Infrared-Visual Fusion for Enhanced Drone Detection, Tracking and Payload Identification in Surveillance Videos” [1]. In this work, we focus specifically on the bird–drone surveillance portion of the data. The payload-identification subset is outside the scope of AVID. This narrower focus allows us to study the setting in which distant aerial targets are observed simultaneously in visible and thermal imagery.

The paired dataset considered in this paper is formed from the source Train and Validation splits, where correspondence between RGB and infrared frames can be established directly from the available records. We retain the original video identities so that the temporal structure of each sequence remains available. We do not include the source Test split in the paired benchmark because the RGB and infrared files use different numbering schemes and the available materials do not provide an authoritative frame-to-frame correspondence. Rather than infer these pairs from filenames or visual similarity, we leave the Test data outside the paired analysis.

3.2 Curation and annotations

The source Train and Validation exports contain two image variants for each frame key: a clean frame and a noise-augmented version of the same observation. We group the files by their common frame key and retain the clean variant for the dataset described in this paper. All counts reported below refer to these deduplicated frames.

For each retained frame, we preserve the original video identifier and frame index. This keeps the connection between individual images and their source sequence intact and allows the paired videos to be reconstructed in their original temporal order.

The bounding-box annotations are inherited from the source challenge data. The verified class mapping contains two categories: bird and drone. The materials available to us do not describe the full annotation process in enough detail to determine, for example, how many annotators labeled each frame or how disagreements were resolved. We therefore use the provided annotations as given and avoid making claims about an annotation procedure that is not documented in the source materials.

Table 2: Composition of the paired AVID corpus. Counts refer to the deduplicated Train and Validation splits for which RGB–infrared correspondence is available. Instance counts correspond to annotated bounding boxes.

Split	RGB–IR pairs	Videos	Bird instances	Drone instances	Total instances
Train	28,790	102	14,113	18,335	32,448
Validation	3,250	13	1,358	2,146	3,504
Total	32,040	115	15,471	20,481	35,952

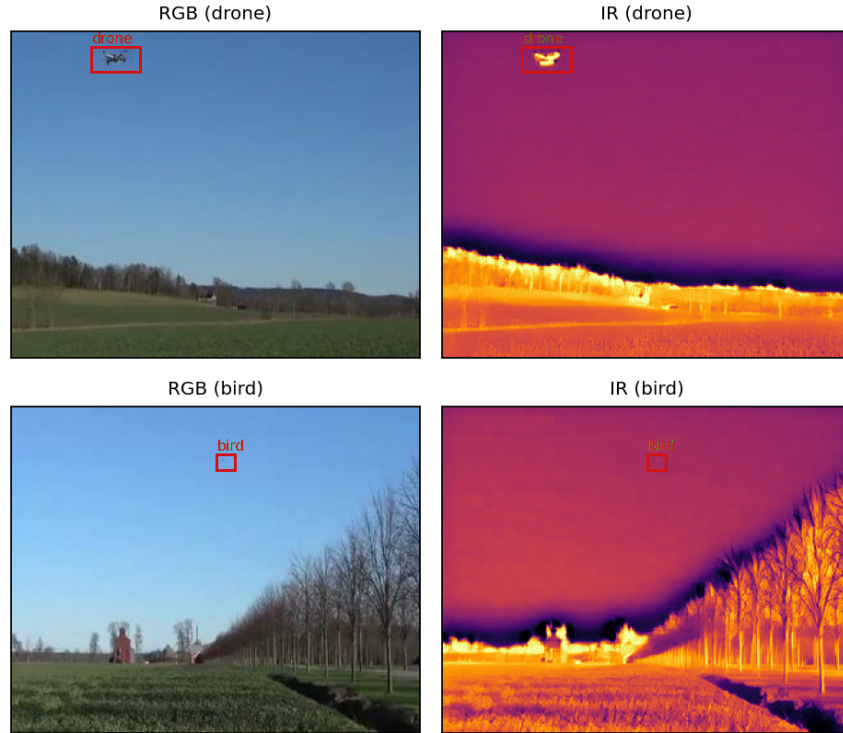


Figure 1: Representative RGB–infrared pairs from AVID. A drone example is shown in the top row and a bird example in the bottom row. Red boxes indicate the target annotations. RGB and infrared observe the same scene and target, but residual spatial displacement can remain between the two modalities.

3.3 Dataset composition

Table 2 summarizes the paired dataset. The Train split contains 28,790 RGB–infrared frame pairs from 102 videos, while the Validation split contains 3,250 pairs from 13 videos. Together, the dataset contains 32,040 paired frames from 115 short surveillance videos.

The 115 videos consist of 76 drone sequences and 39 bird sequences. Individual clips contain between 50 and 329 paired frames, with a median length of 306 frames and a mean of 278.6 frames. Across all frames, the dataset contains 35,952 annotated target instances: 15,471 birds and 20,481 drones.

Figure 1 shows representative RGB–infrared pairs from the dataset. Two properties are apparent even from these examples. First, the aerial targets occupy only a small fraction of the full image. Second, the appearance of the same target can differ substantially between the visible and thermal views. The paired frames should therefore be understood as observations of the same scene from two sensors, rather than as perfectly registered pixel-to-pixel measurements.

4 The geometry of the dataset

4.1 A concentrated tiny-object regime

The defining geometric property of AVID is the small size of its targets. Under the COCO area convention [26], 35,581 of the 35,952 annotated instances, or 99.0%, fall within the “small” category. Only 371 instances are “medium,” and none are “large.”

This conventional partition therefore provides little resolution within the range that matters most for this dataset.

To describe target size more precisely, we use the geometric-mean side length

$$s = \sqrt{w_{\text{px}} h_{\text{px}}}, \quad (4)$$

where w_{px} and h_{px} are the width and height of the bounding box in pixels. This quantity gives a single length scale while preserving the area of the original box.

Figure 2a shows the resulting distribution. The concentration around very small targets is pronounced. The 12–16 pixel interval alone contains 17,706 instances, and the combined 8–16 pixel range contains 24,489 instances, or 68.1% of the entire corpus. Only 103 instances fall below 8 pixels, and no annotated target has a geometric-mean side larger than 48 pixels. We therefore use finer scale intervals throughout the benchmark rather than grouping nearly the entire dataset into a single “small” category.

This scale distribution is important for interpreting the paired RGB–infrared observations. When a target occupies only 10 or 15 pixels, a displacement of a few pixels between the two sensors can correspond to a substantial fraction of the object itself. Alignment therefore needs to be considered together with target size rather than only in image coordinates.

4.2 Weak RGB–infrared alignment

The RGB and infrared streams observe the same scene through separate optical systems. Their fields of view are sufficiently similar to form paired observations, but the two images are not assumed to be registered pixel by pixel. We characterize this residual displacement by estimating a single global translation between each RGB–infrared pair using normalized mutual information (NMI) computed on grayscale images. Horizontal and vertical translations are searched over integer offsets in $[-12, 12]$ pixels.

We measure this displacement on a sample of 300 paired frames. Across the sample, the estimated translation magnitude has a mean of 4.66 pixels, a median of 1.0 pixel, and a standard deviation of 6.44 pixels. Forty-six percent of the pairs have an estimated integer shift of zero, while 36.0% exceed 2 pixels and 32.3% exceed 5 pixels (Figure 2b).

The finite search range places an important limit on how these values should be interpreted. Because each image axis is searched only within ± 12 pixels, the largest translation magnitude that can be represented is

$$\sqrt{12^2 + 12^2} = 16.97 \text{ pixels}. \quad (5)$$

Pairs near this boundary may have larger true offsets. We therefore treat the upper end of the measured distribution as censored rather than interpreting 16.97 pixels as a physical maximum of the sensor displacement.

Pixel displacement alone, however, does not capture the difficulty of the alignment problem. A five-pixel offset has very different consequences for a 10-pixel target and a 50-pixel target. We therefore relate the estimated displacement to the target scale in the same frame using the object-relative ratio defined in Eq. 1.

Across the 300 sampled frames, the median target side is 14.07 pixels. The median object-relative displacement is $\rho = 0.061$, while the mean is 0.339. The upper part of the distribution is more informative than either average. In particular, 33.0% of the sampled pairs have $\rho > 0.25$, 29.7% have $\rho > 0.5$, and 16.7% have $\rho > 1$. In other words, for approximately one sixth of the sampled pairs, the estimated cross-modal displacement is larger than the median target side in that frame.

Figure 2c shows the corresponding alignment burden curve from Eq. 2. Rather than reducing alignment to a single threshold, the curve reports the fraction of pairs whose estimated displacement exceeds any chosen fraction of the target size. The unit-scale burden defined in Eq. 3 is

$$\hat{B}_1 = 0.304. \quad (6)$$

This value can be interpreted as an average displacement of 30.4% of a target side after ratios above one are clipped at one. Together, the burden curve and its scalar summary describe alignment in units that are directly related to the objects being observed.

The alignment ratio and burden curve should be interpreted as summaries of global cross-modal displacement, not as per-object registration ground truth. A single translation cannot capture local parallax, lens distortion, or depth-dependent differences between the two views. The purpose of these quantities is therefore narrower: they describe the scale of the residual RGB–infrared displacement relative to the targets that appear in the dataset.

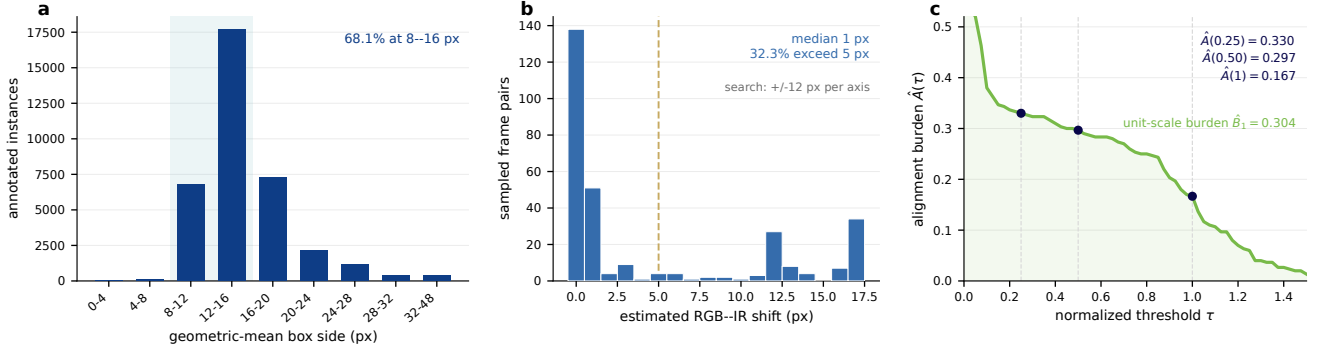


Figure 2: Target scale and weak RGB-infrared alignment in AVID. **a**, Distribution of geometric-mean target side length across the paired Train and Validation data. The shaded 8–16 pixel range contains 68.1% of all annotated instances. **b**, Estimated global RGB-infrared translation magnitude for 300 paired frames. The search is limited to ± 12 pixels along each image axis, so values near the upper end of the distribution are censored by the search range. **c**, Object-relative alignment burden curve $\hat{A}(\tau) = \Pr(\rho > \tau)$. In the measured sample, 33.0%, 29.7%, and 16.7% of pairs exceed one quarter, one half, and one full target side, respectively. The area under the curve up to $\tau = 1$ gives a unit-scale burden of $\hat{\beta}_1 = 0.304$.

5 AVID-C: a controlled perturbation benchmark

5.1 Perturbation families and severity levels

Real multimodal imagery is affected by more than geometric misalignment. Noise, blur, illumination changes, and camera motion can alter the information available in one sensor while leaving the other relatively unaffected. Distorted infrared surveillance has previously been studied in the context of drone localization and tracking, including low visibility, uneven illumination, weak contrast, and background distortion [14]. These effects are particularly important in AVID, where many targets occupy only a few pixels and even moderate degradation can remove a substantial fraction of the target structure.

We introduce AVID-C to study these effects in a controlled way. The benchmark contains seven perturbation families, each evaluated at five levels of increasing strength: Gaussian noise, speckle noise, salt-and-pepper noise, Gaussian blur, uneven illumination, motion blur, and camera instability. These families reflect common forms of image degradation that can arise from sensor noise, optics, environmental conditions, target motion, and platform instability.

The perturbation strengths are defined for the native 320×256 imagery used in AVID. We do not directly adopt the severity values used by ImageNet-C [19], because the spatial scale and image characteristics are substantially different. A blur kernel or camera translation that is mild for a large natural image can remove much of the useful structure from an aerial target that is only 10–15 pixels across. The severity levels in AVID-C are therefore chosen with the scale of the underlying imagery in mind.

Table 3 summarizes the benchmark settings.

A central feature of AVID-C is that the two sensing modalities can be perturbed separately. We consider three natural conditions: RGB is perturbed while infrared remains unchanged, infrared is perturbed while RGB remains unchanged, and both modalities are perturbed. This distinction is important because multimodal robustness is not only a question of how performance changes when the input pair is degraded. It also depends on which sensor is affected and whether the second modality retains complementary information.

For this reason, results under AVID-C are most informative when they are reported separately for each perturbation family and affected modality. Averaging all conditions into a single corruption score can obscure cases in which RGB and infrared respond very differently to the same nominal degradation.

5.2 Relating the perturbation scale to clean imagery

The lowest severity level should represent a meaningful change from the clean data. If the added perturbation is smaller than variation already present in the images, the first part of the severity scale becomes difficult to interpret. We therefore compare the lowest Gaussian-noise level in AVID-C with an estimate of the noise already present in clean RGB and infrared frames.

We estimate the clean-image noise level from 200 frames in each modality using the no-reference estimator of Immerkaer [27]. The method uses a Laplacian response to estimate high-frequency noise without requiring a clean reference image. Because fine texture can also contribute to this response, we use the estimate only to establish the approximate scale of the existing image noise rather than as a measurement of the physical sensor noise.

The median full-frame estimate is 0.302 intensity units for RGB and 0.638 for infrared on the 0–255 intensity scale. By

Table 3: AVID-C perturbation families and severity levels. Each row lists the parameter values used for severity levels 1–5. Noise standard deviations are expressed in 8-bit intensity units. Camera instability combines image translation and rotation, together with a small scale variation.

Family	Parameter	Severity levels 1–5
Gaussian noise	standard deviation	5, 10, 18, 28, 40
Speckle noise	multiplicative σ	0.05, 0.10, 0.18, 0.28, 0.40
Salt-and-pepper	corrupted pixel fraction	0.005, 0.01, 0.02, 0.04, 0.08
Gaussian blur	Gaussian σ (px)	0.5, 1.0, 2.0, 3.5, 5.5
Uneven illumination	field strength	0.10, 0.20, 0.35, 0.50, 0.70
Motion blur	kernel length (px)	3, 7, 11, 17, 25
Camera instability	maximum shift / rotation	(1 px, 0.5°), (3, 1.5°), (6, 3°), (10, 5°), (15, 8°)

comparison, severity 1 of Gaussian noise in AVID-C has a standard deviation of 5 intensity units. This is approximately 16.6 times the median RGB estimate and 7.8 times the median infrared estimate. When the estimate is restricted to the upper third of the image, where many frames contain relatively uniform sky, the corresponding medians are 0.027 for RGB and 0.047 for infrared.

These measurements place the beginning of the AVID-C severity range clearly above the estimated noise floor of the clean imagery. The benchmark therefore starts from a perturbation that produces a measurable change in the input and then increases its strength over the remaining four severity levels. The comparison is shown in Figure 5i.

6 A fixed-representation analysis of the dataset

6.1 Measuring class structure under controlled variation

Target scale, image degradation, and cross-modal alignment change the visual evidence available in AVID. To study these effects independently of a particular trained model, we use a fixed low-dimensional representation of the annotated bird and drone regions. The goal is to measure how the separation between the two classes changes as the underlying data condition changes.

For each annotated instance, we extract a square crop centered on the bounding box with $1.5\times$ padding. The crop is converted to grayscale, resized to 24×24 , and flattened. RGB and infrared patches are projected separately using PCA fitted on the clean data. We retain eight components from each modality, giving an 8-dimensional RGB representation, an 8-dimensional infrared representation, and a 16-dimensional paired representation formed by concatenating the two.

The PCA basis is kept fixed when target scale, corruption severity, or cross-modal displacement is varied. Changes in the measurements therefore reflect changes in the input data under the same representation rather than changes introduced by fitting a new feature space for each condition. Figure 3 summarizes the analysis.

6.2 Class overlap in the fixed representation

Let Z denote the feature representation and let $Y \in \{0, 1\}$ denote the bird or drone class. For each class, we fit a Gaussian distribution

$$Z \mid Y = y \sim \mathcal{N}(\mu_y, \Sigma_y). \quad (7)$$

For two fitted class-conditional Gaussians, the Bhattacharyya distance is

$$D_B = \frac{1}{8}(\mu_1 - \mu_0)^\top \bar{\Sigma}^{-1}(\mu_1 - \mu_0) \quad (8)$$

$$+ \frac{1}{2} \log \frac{\det \bar{\Sigma}}{\sqrt{\det \Sigma_0 \det \Sigma_1}}, \quad \bar{\Sigma} = \frac{\Sigma_0 + \Sigma_1}{2}. \quad (9)$$

Under class priors π_0 and π_1 , the corresponding Bhattacharyya upper bound on the classification error is

$$P_e \leq \sqrt{\pi_0 \pi_1} e^{-D_B}. \quad (10)$$

We use equal class priors throughout the analysis and denote the estimated bound by $\hat{B}$. A smaller value indicates greater separation between the fitted bird and drone distributions in the chosen feature space, while a larger value indicates greater overlap.

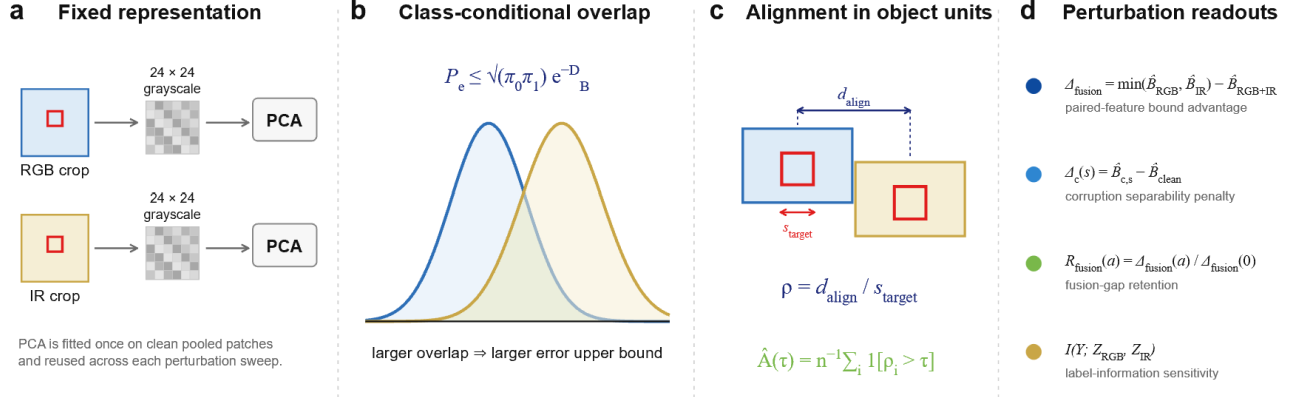


Figure 3: Fixed-representation analysis used to characterize AVID. **a**, Annotated RGB and infrared target crops are converted to 24×24 grayscale patches and projected into PCA spaces fitted on clean data. **b**, Bird and drone class-conditional distributions are compared using Bhattacharyya and Chernoff error upper bounds, with a kernel-density estimate used as a non-Gaussian comparison. **c**, Cross-modal displacement is expressed relative to target size using $\rho = d_{\text{align}} / s_{\text{target}}$. **d**, The resulting measurements summarize the paired-modality advantage, the effect of corruption, the retention of that advantage under increasing shift, and changes in label information.

The Bhattacharyya bound uses a fixed weighting between the two class distributions. We therefore also compute an optimized Chernoff bound, which allows this weighting to vary over $s \in (0, 1)$ and selects the smallest resulting upper bound. At the dominant 12–16 pixel target scale, we additionally compare the result with a kernel-density estimate of class overlap. The KDE calculation uses the same feature space but does not assume Gaussian class-conditional distributions. Agreement across these measurements provides a stronger indication that an observed trend is not specific to a single distributional assumption.

6.3 Comparing single and paired modalities

We use the fitted error bounds to describe how the paired representation differs from either modality alone. We define the fusion gap as

$$\Delta_{\text{fusion}} = \min(\hat{B}_{\text{RGB}}, \hat{B}_{\text{IR}}) - \hat{B}_{\text{RGB+IR}}. \quad (11)$$

A positive fusion gap means that the paired RGB–infrared representation has a lower estimated error bound than the better of the two single-modality representations. A value near zero indicates little difference between the paired representation and the stronger individual modality, while a negative value indicates that the paired representation does not improve the bound.

Because the paired feature contains components from both sensors, the magnitude of Δ_{fusion} depends on the representation as well as on the data. We therefore interpret it as a measure of paired-modality separation within the fixed feature space used in this paper rather than as a model-independent measure of multimodal benefit.

6.4 Effect of corruption

For corruption family c at severity s , we measure the change from the corresponding clean condition as

$$\Delta_c(s) = \hat{B}_{c,s} - \hat{B}_{\text{clean}}. \quad (12)$$

Positive values indicate that the bird and drone distributions become less separated after the perturbation, while negative values indicate increased separation in the fixed representation. The latter can occur, for example, when a transformation removes variation that is not useful for distinguishing the two classes.

This quantity allows corruption effects to be compared relative to the clean condition for each modality. It is particularly useful for AVID-C, where the same nominal perturbation can affect RGB and infrared imagery differently.

6.5 Sensitivity to cross-modal shift

We also examine how the paired representation changes as the infrared target patch is displaced relative to RGB. Let a denote the maximum horizontal and vertical shift used in the controlled alignment experiment. We summarize the remaining paired-modality advantage using the fusion-gap retention

$$R_{\text{fusion}}(a) = \frac{\Delta_{\text{fusion}}(a)}{\Delta_{\text{fusion}}(0)}. \quad (13)$$

A value of $R_{\text{fusion}}(a) = 1$ indicates that the fusion gap is unchanged from the aligned condition, while values below one indicate that part of the initial paired-modality advantage has been lost as the displacement increases.

As a second view of the same shift experiment, we estimate the mutual information between the class label and the joint feature,

$$I(Y; Z_{\text{RGB}}, Z_{\text{IR}}). \quad (14)$$

The class-conditional feature distributions are modeled as Gaussians, which gives analytic conditional entropies, while the entropy of the resulting Gaussian mixture is estimated numerically. We use this quantity to track how much label-related structure remains in the joint representation as cross-modal displacement increases.

Together, these measurements provide complementary views of the same dataset properties. The Bhattacharyya and Chernoff bounds describe class overlap, the fusion gap compares paired and single-modality representations, the corruption penalty measures movement away from the clean condition, and fusion-gap retention and mutual information describe how the paired representation changes as alignment deteriorates.

7 Dataset behavior under scale, corruption, and alignment

7.1 Target scale shapes class separation

Target scale strongly affects the amount of bird–drone structure preserved in the fixed representation. Figure 4a shows the Bhattacharyya error upper bound across the native target-size intervals. Lower values indicate greater separation between the fitted bird and drone distributions.

Across the scale intervals with sufficient examples from both classes, the paired RGB–infrared representation gives the lowest bound. The effect is clearest in the 12–16 pixel interval, which contains the largest fraction of targets in AVID. At this scale, the bounds are $\hat{B}_{\text{RGB}} = 0.148$, $\hat{B}_{\text{IR}} = 0.198$, and $\hat{B}_{\text{RGB+IR}} = 0.046$. The corresponding fusion gap is 0.103 (Figure 4b). Thus, within this representation, combining the two modalities produces substantially greater bird–drone separation than either modality alone at the dominant target scale.

The same ordering is observed with the alternative overlap estimates in Figure 4c. At 12–16 pixels, the optimized Chernoff bounds are 0.131 for RGB, 0.194 for infrared, and 0.040 for the paired representation. The KDE estimates are 0.165, 0.177, and 0.047, respectively. Although the absolute values differ, all three measurements show the same relationship between RGB, infrared, and the paired representation at this scale.

Part of the paired result can be explained by representation dimension. When eight components are retained from each modality, the paired feature is 16-dimensional, whereas each individual modality is represented by eight dimensions. Figure 4d shows that the fitted bounds generally decrease as more PCA components are retained.

We therefore also compare the modalities at equal total dimension. An 8-dimensional paired representation formed from four RGB and four infrared components gives a bound of 0.140 at 12–16 pixels. The corresponding 8-dimensional bounds are 0.148 for RGB and 0.198 for infrared. The paired representation remains lower, but the difference is much smaller than for the 16-dimensional 8+8 representation. Moreover, this ordering is not consistent across all of the larger scale intervals. The fusion gap should therefore be interpreted within the feature representation in which it is measured.

A second control separates the effect of added dimension from the effect of adding a second sensing modality. Repeating the same 8-dimensional RGB vector to produce a redundant 16-dimensional feature changes the fitted bound from 0.148 to 0.127; the corresponding infrared value changes from 0.198 to 0.191. These shifts show that feature dimension alone can change the absolute bound. Even so, both redundant 16-dimensional controls remain well above the 0.046 bound obtained from the 8+8 RGB–infrared representation. The result therefore reflects more than dimension alone, while remaining specific to the representation used here.

7.2 RGB and infrared respond differently to corruption

The seven AVID-C families produce markedly different responses in RGB and infrared imagery. Figure 5a–g shows the Bhattacharyya bound from the clean condition through severity 5 for each perturbation family. The same 1,500 object instances are used across conditions, and the PCA representation remains fixed.

Before perturbation, the bounds are 0.148 for RGB and 0.162 for infrared. At severity 5, the RGB bounds are 0.329 for Gaussian noise, 0.375 for speckle noise, 0.321 for salt-and-pepper noise, 0.186 for Gaussian blur, 0.304 for uneven

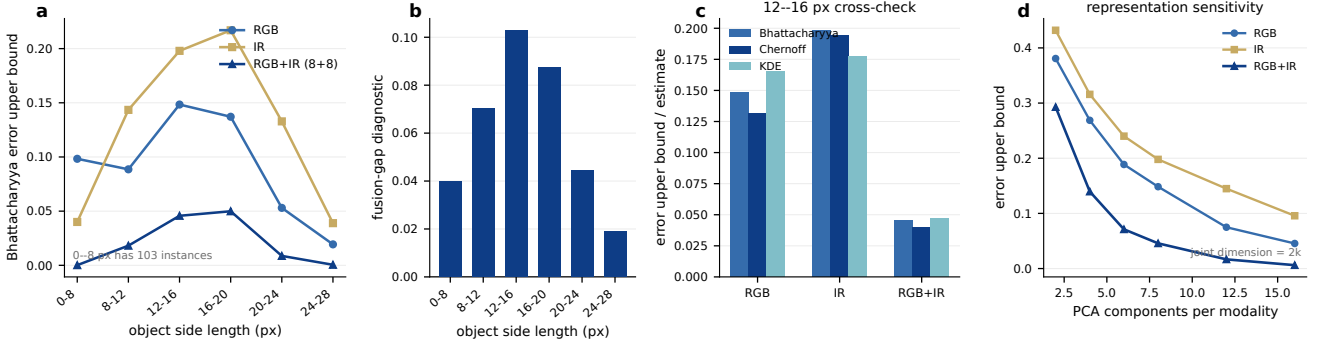


Figure 4: Target scale and bird–drone separation in the fixed representation. **a**, Bhattacharyya error upper bound across native target-size intervals for RGB, infrared, and the 8+8-dimensional paired representation. The 0–8 pixel interval contains only 103 instances and should be interpreted cautiously. **b**, Fusion gap from Eq. 11 across target scales. **c**, Bhattacharyya, optimized Chernoff, and KDE estimates for the dominant 12–16 pixel interval. **d**, Effect of PCA dimension at 12–16 pixels. The paired representation uses k components from each modality and therefore has total dimension $2k$.

illumination, 0.222 for motion blur, and 0.224 for camera instability. The corresponding infrared values are 0.198, 0.222, 0.189, 0.204, 0.245, 0.215, and 0.224.

The difference is particularly clear for the noise perturbations. At the highest severity, Gaussian, speckle, and salt-and-pepper noise increase class overlap substantially more in RGB than in infrared. Other perturbations produce smaller differences between the modalities, and camera instability reaches the same bound of 0.224 for both at severity 5.

Averaged across the seven families at severity 5, the increase over the clean bound is 0.132 for RGB and 0.052 for infrared (Figure 5h). Expressed relative to the clean condition, the mean severity-5 bound is 1.90 times the RGB baseline and 1.32 times the infrared baseline. These averages summarize the overall trend in this dataset, but the individual curves remain more informative because the effect depends strongly on the type of degradation.

The response to Gaussian blur illustrates an important feature of these measurements. The RGB bound initially decreases under moderate blur before increasing again at stronger severities. One possible explanation is that mild smoothing removes high-frequency variation within each class before stronger blur begins to remove class-specific structure. Regardless of the cause, this behavior shows that increasing the strength of a physical perturbation does not require every statistical measure of class overlap to change monotonically. The full severity curve is therefore more informative than a single endpoint.

7.3 Cross-modal shift weakens paired separation

We next study how the paired representation changes when the two modalities are progressively displaced. The infrared target patch is translated horizontally and vertically within a specified per-axis shift range while the RGB patch remains fixed. The representation is defined at zero added shift and kept unchanged as the displacement increases. For each nonzero shift range, results are averaged across five random draws.

Figure 6 shows the response through a maximum per-axis shift of 12 pixels, matching the search range used to characterize alignment in the real paired frames. At zero added shift, the fusion gap is 0.111. At a 12-pixel shift range, it decreases to 0.093. The corresponding fusion-gap retention is

$$R_{\text{fusion}}(12) = 0.844, \quad (15)$$

meaning that 84.4% of the initial fusion gap remains at the largest shift considered.

The mutual-information estimate shows a similar trend. The estimated label information in the joint RGB–infrared representation decreases from 0.715 nats at zero added shift to 0.647 nats at a 12-pixel shift range, corresponding to a 9.5% reduction. Both measurements therefore indicate a gradual loss of useful joint structure as the two target observations are moved farther apart.

The controlled shift experiment and the alignment burden curve describe different aspects of the dataset. The burden curve in Figure 2c measures how frequently different object-relative displacements occur in the observed RGB–infrared pairs. The experiment in Figure 6, in contrast, measures how the paired feature changes as additional displacement is introduced. The first characterizes the alignment present in the data; the second shows how sensitive the paired representation is to increasing geometric disagreement.

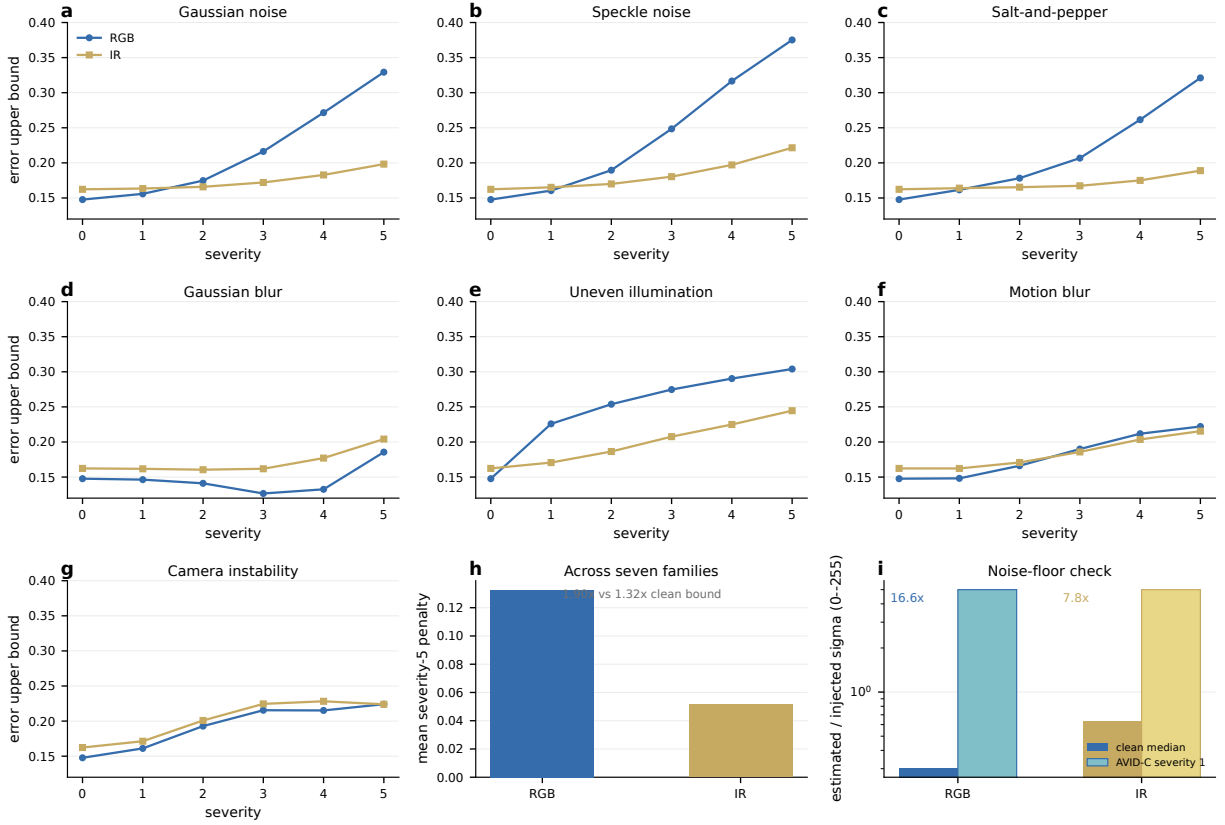


Figure 5: Effect of AVID-C on RGB and infrared class separation. Panels a–g show the Bhattacharyya error upper bound from the clean condition (severity 0) through severity 5 for each perturbation family. h, Mean increase over the clean bound at severity 5 across the seven families. i, Estimated noise level in clean RGB and infrared frames compared with Gaussian-noise severity 1. The injected standard deviation is 16.6 times the median RGB estimate and 7.8 times the median infrared estimate.

7.4 Scale, corruption, and alignment describe different sources of difficulty

The three analyses capture different limitations of tiny-target multimodal sensing. Target scale determines how much spatial structure is available in the first place. Corruption changes the information carried by an individual sensor. Cross-modal displacement changes how well the local evidence from the two sensors corresponds.

These effects are closely connected in AVID because most targets occupy only 8–16 pixels. At this scale, a few pixels of blur or displacement are not small relative to the object. This also explains why alignment expressed in target units is more informative than alignment expressed only in pixels: the relevant question is not simply how far the images are displaced, but how large that displacement is compared with the target being observed.

Taken together, the results show that the difficulty of AVID cannot be described by target category alone. It depends on the scale of the object, the condition of each sensing modality, and the geometric agreement between them. Reporting these factors alongside downstream performance provides a clearer picture of where multimodal methods succeed and where their gains begin to weaken.

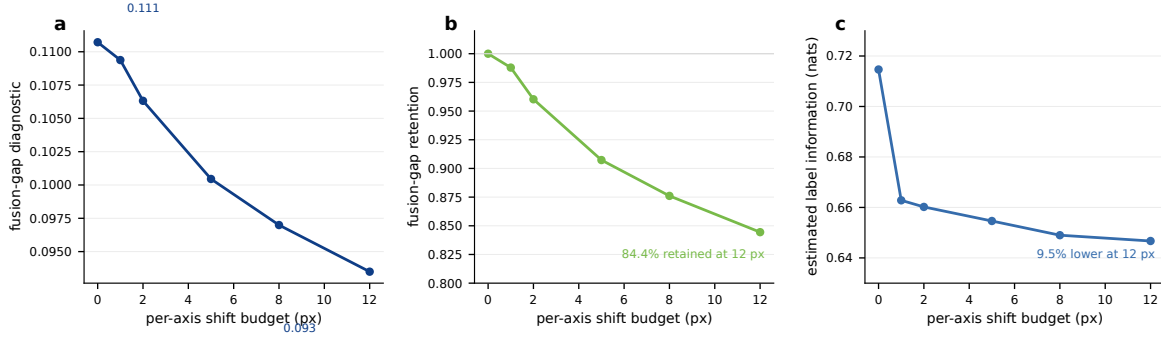


Figure 6: Sensitivity of the paired representation to cross-modal shift. The infrared target patch is translated relative to RGB using random horizontal and vertical offsets within the indicated per-axis range. **a**, Fusion gap as the allowed displacement increases. **b**, Fusion-gap retention relative to the zero-shift condition; 84.4% of the initial gap remains at a 12-pixel shift range. **c**, Estimated label information in the joint representation, which decreases by 9.5% over the same range.

8 Benchmark use and reporting

AVID is intended to support a range of downstream tasks while keeping the evaluated data condition visible. Aggregate task metrics remain useful, but they are easier to interpret when target scale, sensing modality, corruption condition, and cross-modal alignment are reported alongside them. We therefore recommend the following reporting conventions.

8.1 Clean paired evaluation

For evaluation on the unperturbed data, results should state whether the input consists of RGB, infrared, or paired RGB–infrared observations. Any resizing or cropping applied to the native 320×256 frames should also be reported, since these operations change the apparent target scale.

Results should identify the data split and the number of frames or videos used. The paired benchmark in this paper is defined only on the source Train and Validation data, where RGB–infrared correspondence can be established. The source Test data should not be treated as a paired split unless an authoritative correspondence between its RGB and infrared frames becomes available.

8.2 Scale-aware evaluation

Object-level results are most informative when they are stratified by the geometric-mean target side

$$s = \sqrt{w_{\text{px}} h_{\text{px}}}. \quad (16)$$

For AVID, we use the intervals 0–8, 8–12, 12–16, 16–20, 20–24, 24–28, 28–32, and 32–48 pixels. The number of examples in each interval should be shown together with the corresponding task metric. This is particularly important for the 0–8 pixel range, which contains only 103 annotated instances.

For sequence-level tasks, the same idea can be applied using a frame-level scale distribution or a sequence statistic such as the median target side. The purpose is to make clear whether a result is dominated by the 8–16 pixel regime that accounts for most of the dataset.

8.3 Corruption evaluation

Results under AVID-C should identify the perturbation family, severity level, and affected modality. We distinguish three settings:

1. RGB perturbed while infrared remains unchanged,
2. infrared perturbed while RGB remains unchanged, and
3. both modalities perturbed.

When both streams are perturbed, the relationship between their random perturbations should also be stated. Family-level results are preferable to reporting only a single corruption average because the two sensing modalities can respond very differently to the same type of degradation.

Table 4: Recommended reporting items for AVID. The aim is to make the evaluated data condition clear alongside any downstream metric.

Item	What to report
Input	RGB, infrared, or paired RGB-IR; native or resized resolution
Data subset	split, video identities when relevant, and frame counts
Target scale	geometric-mean side intervals and number of instances
Corruption	family, severity, and affected modality or modalities
Alignment	estimator, search range, and ρ thresholds or burden curve
Aggregation	per-scale and per-family values in addition to any overall average
Fixed representation	crop definition, PCA dimension, class priors, and density model when comparing with Section 6

The full severity curve should be shown whenever possible. As observed in the blur experiments, a stronger physical perturbation does not require every statistical measure or task metric to vary monotonically.

8.4 Alignment-aware evaluation

When an alignment estimate is available, raw pixel displacement should be accompanied by a target-relative quantity. For the alignment ratio

$$\rho = \frac{d_{\text{align}}}{s_{\text{target}}}, \quad (17)$$

useful reference thresholds are $\tau \in \{0.25, 0.5, 1.0\}$. These correspond to estimated displacements larger than one quarter, one half, and one full target side.

A compact alignment summary can therefore report

$$\hat{A}(0.25), \quad \hat{A}(0.5), \quad \hat{A}(1.0), \quad (18)$$

together with the unit-scale burden $\hat{B}_1$.

The alignment estimator and its search range should be stated because both affect the resulting distribution. If a different registration method is used, the object-relative quantities should be recomputed from that estimator rather than compared directly with values obtained under a different displacement definition.

One advantage of the burden curve is that it reduces the direct dependence on image resolution. Across datasets with annotated targets, $\hat{A}(\tau)$ asks the same basic question: what fraction of paired observations have an estimated cross-modal displacement larger than τ target sides? Differences in optics and registration methods remain, but the result is easier to interpret than pixel displacement alone.

8.5 Comparison with the fixed-representation analysis

The numerical values reported in Section 6 depend on the representation used in that analysis. Direct numerical comparison therefore requires the same crop definition, 24×24 grayscale representation, PCA dimensions, equal class priors, and density models.

Alternative representations are useful and encouraged, but they define a different measurement space. Their absolute Bhattacharyya, Chernoff, fusion-gap, or mutual-information values should therefore be interpreted within that representation rather than treated as interchangeable with those reported here.

9 Data access and licensing

The dataset and benchmark materials are available at https://github.com/rdharini2001/ICIP_2025_Data_Release. The media are distributed under CC-BY-NC-4.0. Updates that change pair correspondence, annotations, or derived metadata should be versioned so that results can be associated with a specific state of the dataset.

10 Responsible use

AVID contains surveillance imagery of birds and unmanned aerial vehicles. Relevant research settings include multimodal perception, tiny-target recognition, sensor robustness, weak cross-modal alignment, tracking, and counter-UAV sensing. The

dataset does not contain person-identity labels or personal-attribute annotations; its unit of analysis is the distant aerial target and the paired sensor observation.

Counter-UAV sensing is inherently dual use. The same perception technology that can support airspace safety and infrastructure protection can also become part of broader surveillance or interdiction systems. Results on AVID should therefore be interpreted as measurements of perception under the conditions represented in the benchmark, not as evidence that a system is safe or appropriate for autonomous operational use. Deployment requires separate consideration of legal constraints, false-alarm costs, human oversight, and the consequences of incorrect decisions.

Birds are treated as a first-class target category rather than as background clutter. This distinction is important in long-range surveillance because birds can be visually similar to small drones when only a few pixels describe the target. Excluding them would remove a central source of ambiguity in the sensing problem that AVID is intended to represent.

11 Limitations

The paired benchmark does not include the source Test split. The available RGB and infrared Test files cannot be paired reliably from their identifiers, and the source materials do not provide an authoritative correspondence. We therefore restrict paired analysis to the Train and Validation data.

The original annotation procedure is only partially documented. The bounding boxes and class labels are inherited from the source challenge data. The available materials do not specify the complete annotator assignment, review process, or adjudication procedure. As a result, this paper cannot quantify annotation disagreement or label uncertainty beyond the annotations themselves.

Alignment is measured on a subset of the paired data. The alignment statistics are based on 300 frame pairs rather than all 32,040 pairs. They characterize the measured subset and should not be interpreted as frame-level registration metadata for the complete dataset.

The alignment model is limited to global translation. The NMI measurement estimates one horizontal and vertical translation for each image pair. It cannot describe local parallax, lens distortion, depth-dependent displacement, or other spatially varying differences between the two sensors.

The upper end of the alignment distribution is censored. The search is limited to ± 12 pixels along each image axis. A displacement that reaches this boundary may in reality be larger. Consequently, the largest measured values and the upper end of the burden curve should not be interpreted as complete physical calibration of the sensor pair.

The class-separation analysis depends on the chosen representation. The main analysis uses eight PCA components for RGB and eight for infrared together with Gaussian class-conditional models. The paired representation therefore has greater total dimensionality than either single modality. We include equal-dimension and redundant-dimension comparisons to show how this choice affects the bound. The KDE analysis relaxes the Gaussian assumption but still operates in the same low-dimensional feature space.

The fixed-representation analysis assumes the target location is known. The analysis uses crops centered on ground-truth bounding boxes. It therefore characterizes the visual evidence available once the target location has been supplied. It does not measure full-frame detection, localization, temporal association, or the difficulty of finding a tiny object in a large scene.

AVID-C covers a selected set of perturbations. The seven families represent common forms of noise, blur, illumination change, motion, and camera instability, but they do not span every possible source of degradation. Thermal artifacts, atmospheric effects, compression failures, lens contamination, rolling shutter, and more complex platform motion are outside the present benchmark.

Perturbation severity is not equivalent to task difficulty. Severity in AVID-C describes the magnitude of the image transformation. A larger transformation need not produce a monotonic change in every representation or downstream metric. The non-monotonic blur response provides one example of this distinction.

Capture diversity remains finite. The 115 videos contain varied backgrounds, target appearances, and thermal conditions, but they cannot cover every geography, weather condition, sensor design, drone airframe, bird species, viewing distance, or observation geometry. Evaluation on additional datasets remains necessary for claims about broader generalization.

12 Conclusion

We introduced AVID, a paired RGB–infrared dataset and benchmark for tiny aerial targets. The dataset contains 32,040 paired frames from 115 surveillance videos with 35,952 annotated bird and drone instances. Its defining characteristics are a strong concentration of very small targets, measurable disagreement between the RGB and infrared views, and modality-dependent sensitivity to image degradation.

A central observation of the dataset is that alignment should be interpreted relative to target size. The object-relative ratio $\rho = d_{\text{align}}/s_{\text{target}}$ expresses displacement in units of the object being observed, while the alignment burden curve describes how often a chosen fraction of the target side is exceeded. In the 300-pair alignment sample, 16.7% of pairs have an estimated displacement larger than one median target side, and the unit-scale burden is 0.304.

AVID-C extends this view to sensor degradation through seven perturbation families and five severity levels. The results show that RGB and infrared do not respond uniformly to the same degradation and that the relative behavior depends strongly on the perturbation family. The fixed-representation analysis further shows that paired RGB–infrared information can provide greater bird–drone separation at the dominant 12–16 pixel scale, while the representation controls make clear that the magnitude of this effect depends on the feature space in which it is measured.

The broader value of AVID lies in making these data properties visible. Tiny-target scale, cross-modal displacement, and sensor degradation are not secondary details in this setting; they shape the information available to any method operating on the data. Reporting them alongside downstream performance should make comparisons on AVID easier to interpret and should help separate improvements in modeling from changes in the difficulty of the evaluated data.

Acknowledgements

This work builds on the dataset-collection effort for the IEEE ICIP 2025 Video and Image Processing Cup, “Infrared-Visual Fusion for Enhanced Drone Detection, Tracking and Payload Identification in Surveillance Videos.” We thank the organizers and contributors who made the source imagery available for research. The present work focuses on the paired bird–drone surveillance data and does not include the payload-identification component of the original challenge.

References

- [1] IEEE Signal Processing Society. VIP Cup 2025 at ICIP 2025: Infrared-Visual Fusion for Enhanced Drone Detection, Tracking and Payload Identification in Surveillance Videos. <https://signalprocessingsociety.org/community-involvement/vip-cup-2025-icip-2025>, 2025. Accessed September 2026.
- [2] Soonmin Hwang, Jaesik Park, Namil Kim, Yookyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baselines. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1037–1045, 2015.
- [3] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. LLVIP: A visible-infrared paired dataset for low-light vision. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3489–3497, 2021.
- [4] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [5] Teledyne FLIR. FLIR thermal dataset for algorithm training. <https://oem.flir.com/solutions/automotive/adas-dataset-form/>, 2018.
- [6] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2022.
- [7] Tao Peng, Qiang Li, and Pengfei Zhu. RGB-T crowd counting from drone: A benchmark and MMCCN network. In *Asian Conference on Computer Vision (ACCV)*, 2020.

- [8] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [9] Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Guodong Guo, Qixiang Ye, Jianbin Jiao, Jian Zhao, and Zhenjun Han. Anti-UAV: A large-scale benchmark for vision-based UAV tracking. *IEEE Transactions on Multimedia*, 25:486–500, 2023.
- [10] Jian Zhao and collaborators. Anti-UAV official repository and dataset releases. <https://github.com/ZhaoJ9014/Anti-UAV>, 2026. Accessed September 2026.
- [11] S. Sethu Selvi, S. Pavithraa, R. Dharini, and E. Chaitra. A deep learning approach to classify drones and birds. In *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, pages 1–5, 2022. doi: 10.1109/MysuruCon55714.2022.9972589.
- [12] S Sethu Selvi, Vishnu R Rao, Dharini Raghavan, and Shefali Singh. Modalfusenet architecture for drone localization. In *2026 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pages 1–6, 2026. doi: 10.1109/CONECCT68978.2026.11700652.
- [13] Dharini Raghavan and S. Sethu Selvi. SetNet: A sparse ensemble network for drone localization and zero shot drone tracking in real time surveillance videos. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5, 2023. doi: 10.1109/ICASSPW59220.2023.10193217.
- [14] Dharini Raghavan, Sethu Selvi S, Raghuram S, Vishnu R. Rao, and Srinidhi A. V. IR-SetNet: Sparsity aware ensemble network for infrared imaging based drone localization and tracking in distorted surveillance videos. In *2023 IEEE International Conference on Image Processing Challenges and Workshops (ICIPCW)*, 2023. doi: 10.1109/ICIPCW59416.2023.10328384.
- [15] Yang Liu, Zhihao Sun, Lele Xi, Lele Zhang, Wei Dong, Chen Chen, Maobin Lu, Hailing Fu, and Fang Deng. MMFW-UAV dataset: multi-sensor and multi-view fixed-wing UAV dataset for air-to-air vision tasks. *Scientific Data*, 12:185, 2025. doi: 10.1038/s41597-025-04482-2.
- [16] Tianyang Xu, Jinjie Gu, Xuefeng Zhu, Xiaojun Wu, and Josef Kittler. A tri-modal dataset and a baseline system for tracking unmanned aerial vehicles. *arXiv preprint arXiv:2511.18344*, 2025.
- [17] Yihang Luo, Jun Chen, Chao Xiao, Yingqian Wang, Zhaoxu Li, Qiang Ling, Xu He, Nuo Chen, Gaowei Guo, Hongge Li, Miao Li, Longguang Wang, Yulan Guo, Li Liu, Wei An, and Zhijie Chen. Towards UAV detection in the real world: A new multispectral dataset UAVNet-MS and a new method. *arXiv preprint arXiv:2605.20963*, 2026.
- [18] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 40468–40478, 2026.
- [19] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.
- [20] Xiaoqiong Liu, Yunhe Feng, Shu Hu, Xiaohui Yuan, and Heng Fan. Benchmarking the robustness of UAV tracking against common corruptions. In *IEEE International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2024. doi: 10.1109/MIPR62202.2024.00079.
- [21] Xiaoshuai Hao, Mengchuan Wei, Yifan Yang, Haimei Zhao, Hui Zhang, Yi Zhou, Qiang Wang, Weiming Li, Lingdong Kong, and Jing Zhang. Is your HD map constructor reliable under sensor corruptions? In *Advances in Neural Information Processing Systems*, volume 37, 2024.
- [22] Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlas, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, et al. DataPerf: Benchmarks for data-centric AI development. In *Advances in Neural Information Processing Systems*, volume 36, 2023.
- [23] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. DataComp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

- [24] Jiashuo Liu, Tianyu Wang, Peng Cui, and Hongseok Namkoong. On the need for a language describing distribution shifts: Illustrations on tabular datasets. In *Advances in Neural Information Processing Systems*, volume 36, 2023.
- [25] David Mayo, Jesse Cummings, Xinyu Lin, Dan Gutfreund, Boris Katz, and Andrei Barbu. How hard are computer vision datasets? calibrating dataset difficulty to viewing time. In *Advances in Neural Information Processing Systems*, volume 36, 2023.
- [26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.
- [27] John Immerkær. Fast noise variance estimation. *Computer Vision and Image Understanding*, 64(2):300–302, 1996. doi: 10.1006/cviu.1996.0060.
- [28] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, III Daumé, Hal, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

A Dataset documentation

This appendix follows the structure of Datasheets for Datasets [28] and summarizes the origin, composition, and scope of AVID.

A.1 Motivation

What was the source imagery collected for? The source imagery was collected for the IEEE ICIP 2025 Video and Image Processing Cup on infrared–visual drone surveillance [1].

What part of the source data is included in AVID? AVID focuses on the paired bird–drone surveillance imagery. The payload-identification component of the original challenge is outside the scope of this paper.

A.2 Composition

What does the paired dataset contain? The dataset contains 32,040 RGB–infrared frame pairs from 115 short videos and 35,952 annotated bounding boxes. Train contains 28,790 pairs from 102 videos, while Validation contains 3,250 pairs from 13 videos. The annotations include 15,471 bird instances and 20,481 drone instances.

What is the image resolution? The RGB and infrared frames analyzed in this paper have native resolution 320×256 pixels.

How long are the video sequences? The 115 paired clips contain between 50 and 329 frames. The median sequence length is 306 paired frames and the mean is 278.6.

How small are the targets? Most annotated targets occupy a narrow range of spatial scales. In total, 68.1% have a geometric-mean side between 8 and 16 pixels, and 99.0% fall within the COCO “small” category. Table 5 gives the complete fine-grained histogram.

Table 5: Distribution of target size in AVID. Size is measured by the geometric-mean side length of each bounding box.

Side (px)	0–4	4–8	8–12	12–16	16–20	20–24	24–28	28–32	32–48
Count	0	103	6,783	17,706	7,269	2,155	1,189	376	371

What is not included? The payload-identification data are outside the scope of this paper. The source Test data are also excluded from the paired benchmark because correspondence between their RGB and infrared files cannot be established from the available identifiers.

A.3 Collection and annotation

What viewpoint does the dataset represent? The imagery comes from surveillance cameras observing distant birds and drones in the scene. It should not be confused with aerial-view datasets in which a camera mounted on a UAV looks down at objects on the ground.

Where do the bounding boxes come from? The annotations are inherited from the source challenge data. The full annotation and adjudication procedure is not documented in the materials available for this work, so we do not infer an annotator count or review process that cannot be verified.

Were multiple variants of the same frame present in the source data? Yes. The source Train and Validation exports contained a clean image and a noise-augmented variant associated with the same frame key. AVID uses one clean image for each frame key, and all statistics in this paper refer to those deduplicated frames.

A.4 Access and maintenance

Where is the dataset maintained? The project repository is available at https://github.com/rdharini2001/ICIP_2025_Data_Release.

How should future corrections be handled? Changes to pairing, labels, or derived metadata should be associated with a new dataset version. If an authoritative pairing for the source Test split becomes available, it should be introduced as a new version rather than altering the definition of the current paired corpus.

What license applies to the media? The media are distributed under CC-BY-NC-4.0.

B Alignment measurement

The alignment analysis estimates a global translation between RGB and infrared images for 300 paired frames. RGB is converted to grayscale, and infrared is brought to the same spatial dimensions when required. Candidate integer translations are evaluated over

$$(d_x, d_y) \in [-12, 12]^2. \quad (19)$$

For each candidate displacement, the shifted image pair is compared after removing a 12-pixel border so that translated image boundaries do not enter the measurement. Normalized mutual information is computed from a 32-bin joint histogram:

$$\text{NMI}(A, B) = \frac{H(A) + H(B)}{H(A, B)}. \quad (20)$$

The displacement that maximizes NMI defines the estimated translation. Its Euclidean magnitude is

$$d_{\text{align}} = \sqrt{d_x^2 + d_y^2}. \quad (21)$$

The 300 frames are drawn from the paired Train and Validation data using a fixed random selection. A phase-correlation estimate based on Sobel edge magnitude is also computed as a secondary comparison. Disagreement between the two estimates should not be interpreted as registration ground truth because both methods can behave differently in low-texture regions and under cross-modal intensity changes.

For each sampled frame, the target scale s_i is the median geometric-mean side length of the annotated objects in that frame. The object-relative displacement is then

$$\rho_i = \frac{d_i}{s_i}. \quad (22)$$

The empirical burden curve is

$$\hat{A}(\tau) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}[\rho_i > \tau]. \quad (23)$$

For $T = 1$, the truncated burden score can also be written as

$$\hat{\mathcal{B}}_1 = \int_0^1 \hat{A}(\tau) d\tau = \frac{1}{n} \sum_i \min(\rho_i, 1). \quad (24)$$

This identity gives the score a direct interpretation: it is the average displacement in target-side units after values above one target side are clipped at one.

Table 6: Alignment statistics for the 300-pair sample. The maximum representable radial shift is determined by the ± 12 -pixel search range and should not be interpreted as a physical maximum of the sensor offset.

Quantity	Value
Number of frame pairs	300
Mean estimated shift	4.66 px
Median estimated shift	1.00 px
Standard deviation	6.44 px
Fraction at 0 px	46.0%
Fraction > 2 px	36.0%
Fraction > 5 px	32.3%
Maximum representable radial shift	16.97 px
Median target side	14.07 px
$\hat{A}(0.25)$	0.330
$\hat{A}(0.50)$	0.297
$\hat{A}(1.00)$	0.167
$\hat{\mathcal{B}}_1$	0.304

C Fixed-representation analysis details

C.1 Patch representation and scale analysis

The analysis uses the 35,952 annotated object instances in the paired dataset. For each annotation, we extract a square region centered on the target with $1.5\times$ padding, convert it to grayscale, resize it to 24×24 , and flatten the resulting intensities.

RGB and infrared are projected with separate PCA bases fitted on the clean object patches. The main representation retains eight components from each modality. The paired feature is therefore the 16-dimensional concatenation of the two 8-dimensional vectors. All class-overlap calculations use equal bird and drone priors.

Targets are grouped by their native geometric-mean side length. A scale interval is included only when both bird and drone contribute at least 15 examples. The corresponding Bhattacharyya error upper bounds are shown in Table 7.

Table 7: Bhattacharyya error upper bounds across target scales. The paired representation contains eight PCA components from RGB and eight from infrared. “–” indicates insufficient support under the minimum-count criterion.

	0–8	8–12	12–16	16–20	20–24	24–28	28+
RGB	0.098	0.089	0.148	0.137	0.053	0.019	–
Infrared	0.040	0.143	0.198	0.217	0.133	0.039	–
RGB+IR (8+8)	<0.001	0.018	0.046	0.050	0.009	<0.001	–

The 0–8 pixel interval contains 52 bird instances and 51 drone instances and should therefore be interpreted more cautiously than the larger scale intervals.

C.2 Chernoff and KDE comparison

The Bhattacharyya bound corresponds to the $s = 0.5$ member of the Chernoff family. We also evaluate Chernoff bounds for 99 values of s between 0.01 and 0.99 and retain the smallest value.

At the dominant 12–16 pixel scale, we additionally estimate class overlap using kernel density estimation. The KDE calculation uses 4,000 samples across the two classes and estimates the Bhattacharyya coefficient by importance sampling from an equal mixture of the fitted class densities.

Table 8 shows the three measurements at this scale.

Table 8: Alternative class-overlap estimates at the dominant 12–16 pixel scale.

Representation	Bhattacharyya ($s = 0.5$)	Optimized Chernoff	Optimal s	KDE
RGB	0.1484	0.1314	0.33	0.1650
Infrared	0.1979	0.1944	0.43	0.1774
RGB+IR	0.0457	0.0396	0.37	0.0472

C.3 Representation-dimension controls

The main paired representation uses k components from each modality and therefore has total dimension $2k$. To separate modality complementarity from the effect of dimensionality alone, we consider two additional comparisons.

First, an 8-dimensional paired representation uses four RGB and four infrared components, giving the paired and single-modality features equal total dimension. Second, redundant 16-dimensional controls are formed by repeating the same 8-dimensional RGB or infrared feature without introducing information from a second sensor.

The resulting bounds at 12–16 pixels are shown in Table 9.

These comparisons show that the absolute fusion-gap magnitude depends on the chosen feature dimension. We therefore use it to compare conditions within a fixed representation rather than as a representation-independent measure of multimodal performance.

C.4 Corruption analysis

The corruption analysis uses the same 1,500 annotated object instances across all conditions. PCA is fitted on the clean patches and kept fixed as the perturbation family and severity change.

Table 10 reports the Bhattacharyya bounds at severity 5. The corresponding clean values are 0.1477 for RGB and 0.1623 for infrared.

Table 9: Representation-dimension comparisons at 12–16 pixels. The controls illustrate the sensitivity of the absolute bound to feature dimension.

Representation	Total dimension	Bound	Composition
RGB PCA	8	0.1484	RGB only
Infrared PCA	8	0.1979	infrared only
RGB+IR	8	0.1396	4 RGB + 4 IR
RGB duplicated	16	0.1272	repeated RGB feature
Infrared duplicated	16	0.1913	repeated infrared feature
RGB+IR	16	0.0457	8 RGB + 8 IR

Table 10: Bhattacharyya error upper bounds at severity 5 of AVID-C. Clean values are 0.1477 for RGB and 0.1623 for infrared.

Perturbation	RGB	Infrared
Gaussian noise	0.3292	0.1982
Speckle noise	0.3752	0.2215
Salt-and-pepper	0.3211	0.1889
Gaussian blur	0.1855	0.2041
Uneven illumination	0.3040	0.2446
Motion blur	0.2221	0.2155
Camera instability	0.2241	0.2240

C.5 Controlled cross-modal shift

To study sensitivity to increasing geometric disagreement, the infrared object patch is translated while the RGB patch remains fixed. The shift budget is defined independently along the horizontal and vertical axes.

The tested values are

$$a \in \{1, 2, 5, 8, 12, 17\} \text{ pixels.} \quad (25)$$

For each nonzero value of a , five trials are evaluated. Within a trial, horizontal and vertical translations are drawn independently from $[-a, a]$. Native-image displacement is mapped into the 24×24 canonical patch space using a factor of $24/16 = 1.5$, corresponding to the dominant 12–16 pixel target regime.

The main figure reports shifts through 12 pixels so that its range matches the per-axis search used for the measured alignment distribution. The 17-pixel condition lies beyond that range and is therefore omitted from the main comparison.

Table 11: Controlled cross-modal shift results. The shift value is a per-axis displacement budget rather than an exact radial translation.

Shift budget	Joint upper bound	Fusion gap	Gap retention	MI (nats)
0	0.0370	0.1107	1.000	0.7147
1	0.0383	0.1094	0.988	0.6628
2	0.0414	0.1063	0.960	0.6602
5	0.0472	0.1005	0.907	0.6546
8	0.0507	0.0970	0.876	0.6490
12	0.0542	0.0935	0.844	0.6467

D Full-frame AVID-C examples

Figures 7 and 8 show all seven perturbation families across the five severity levels on representative full-resolution frames. These examples provide a visual reference for the transformation magnitudes listed in Table 3.

AVID-C corruption suite on a real full-resolution RGB frame, severities 1-5

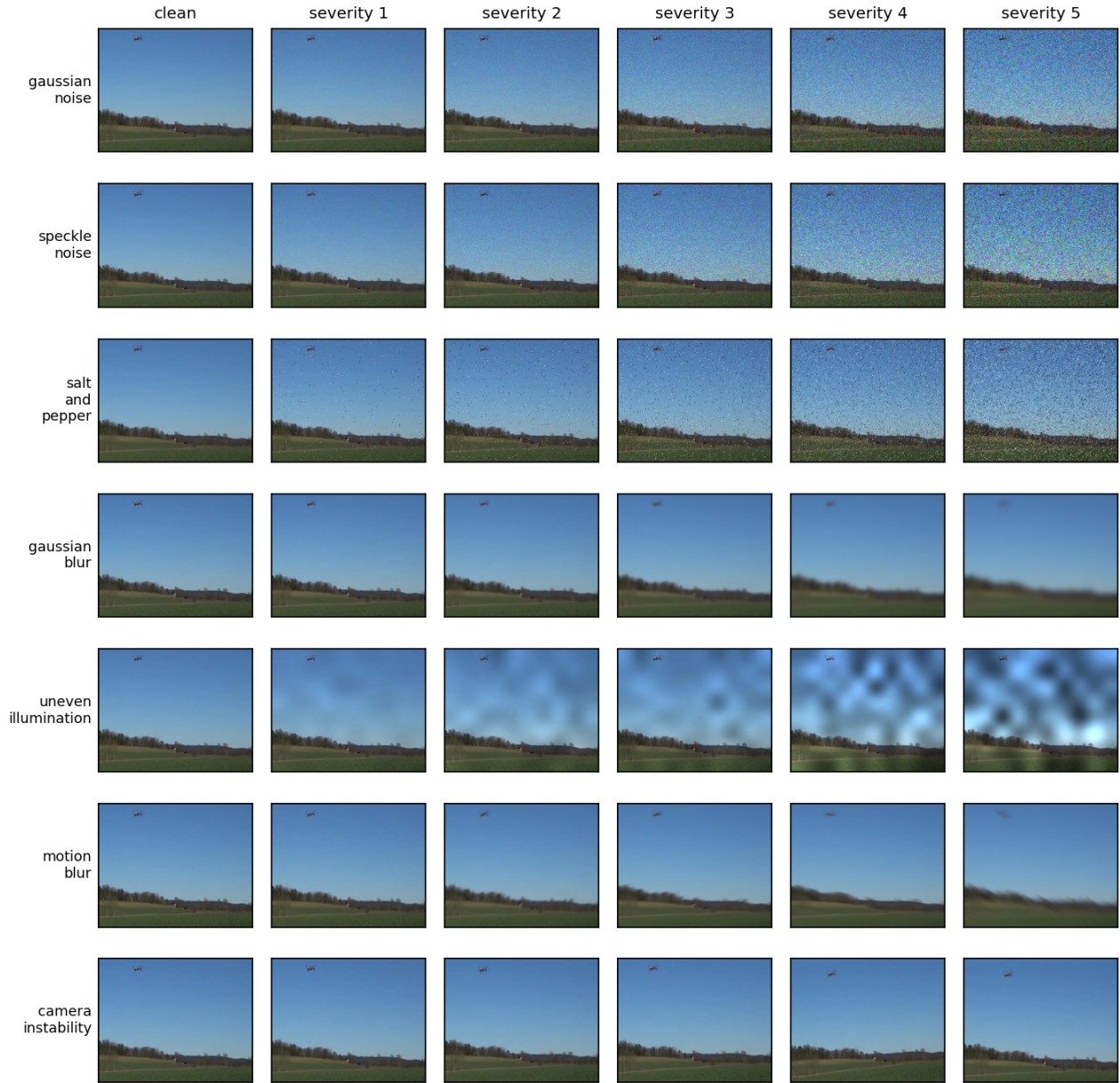


Figure 7: RGB examples from AVID-C. Each row corresponds to one perturbation family, with severity increasing from left to right. Parameter values are given in Table 3.

AVID-C corruption suite on a real full-resolution IR frame, severities 1-5

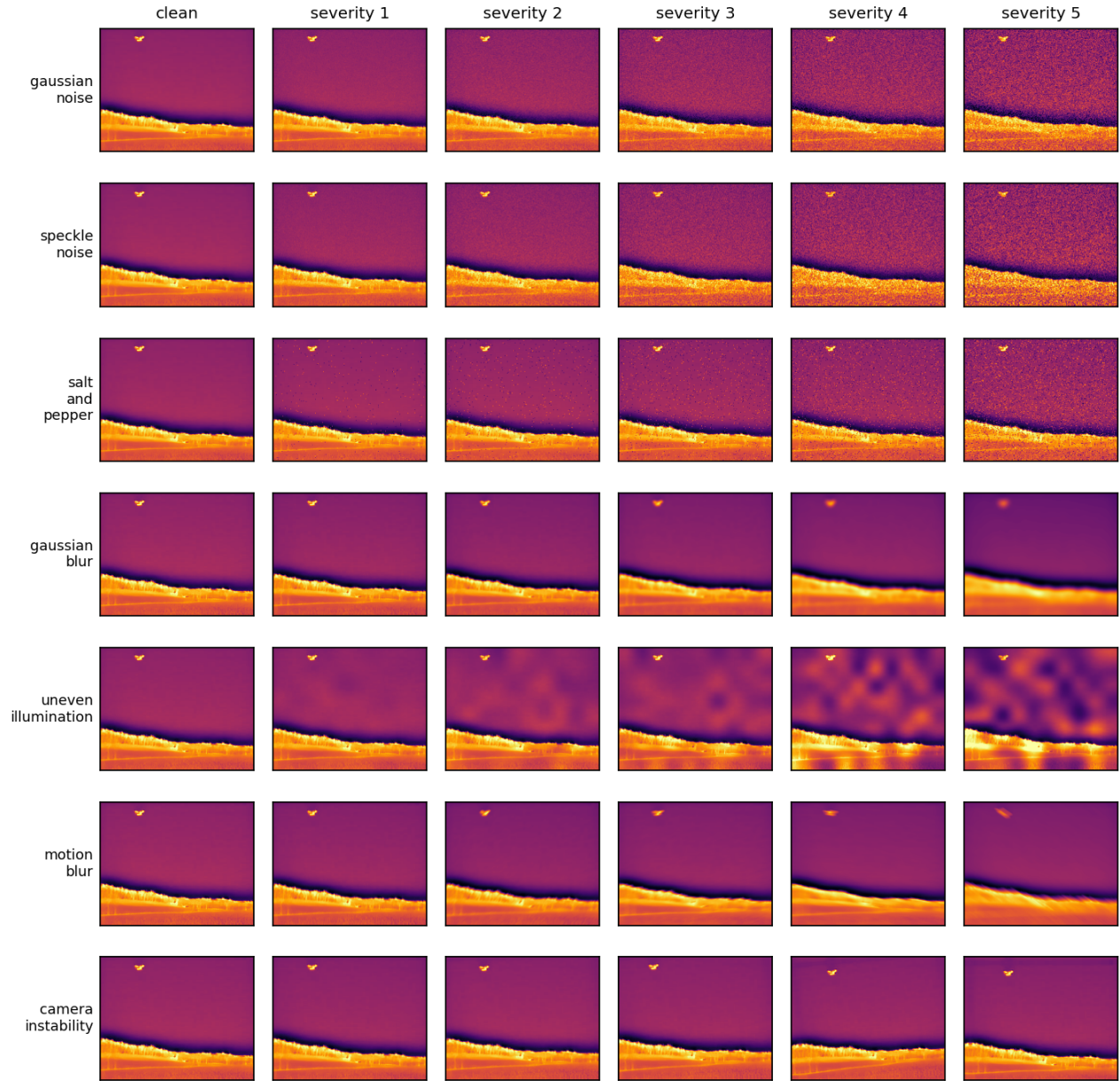


Figure 8: Infrared examples from AVID-C. Each row shows the same perturbation family and severity progression as the RGB examples, applied to the thermal imagery.

E Additional dataset examples

Figure 9 shows additional paired observations drawn from different source videos. The examples illustrate variation in background, target appearance, target count, and thermal contrast across the dataset.

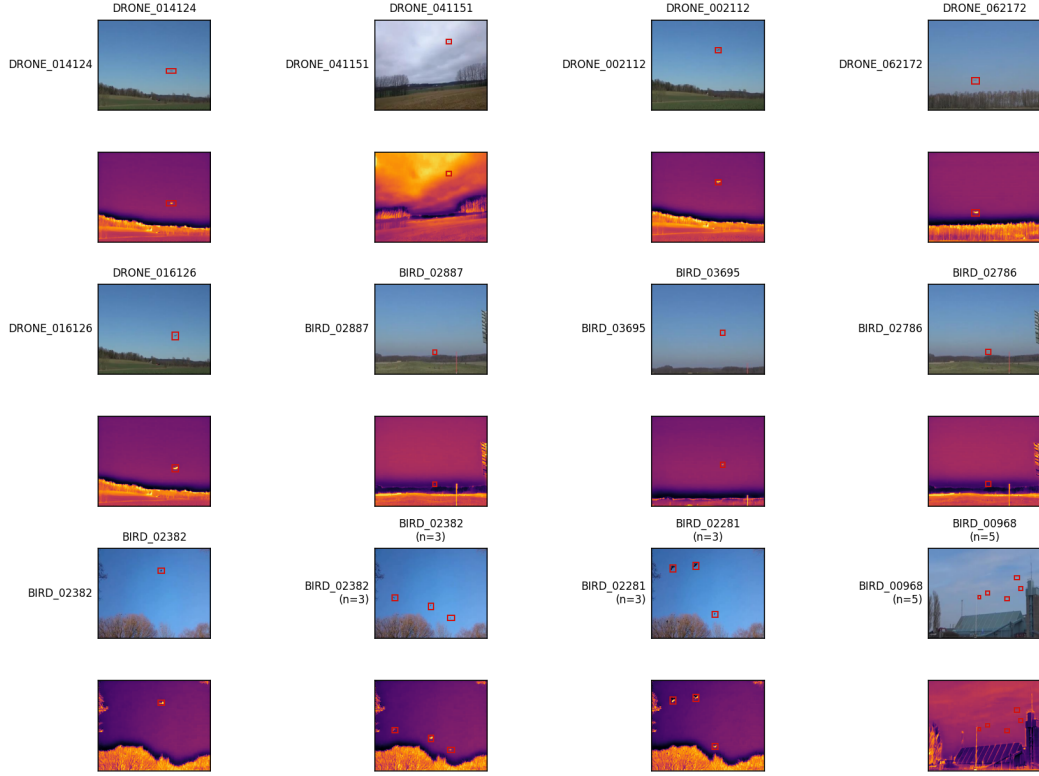


Figure 9: Additional paired examples from AVID. The montage shows scenes from distinct source videos with variation in background, target appearance, target count, and thermal contrast. Red boxes indicate the target annotations.