

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

Dharini Raghavan^{1,*} Vishnu Rao^{2,*} Sethu Selvi S³

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

³Ramaiah Institute of Technology

draghavan7@gatech.edu

 [Code](#)

 [Hugging Face](#)

[Research website](#)

Abstract

Detecting distant drones and distinguishing them from birds is difficult when targets occupy only a few pixels in surveillance videos. Using both RGB and infrared cameras can help, but the two views are not always aligned, and noise, blur, and other image distortions can affect each sensor differently. Understanding these effects is important for developing reliable drone detection systems. For the IEEE ICIP 2025 Video and Image Processing Cup, we curated bird and drone surveillance footage from an earlier multi-sensor dataset. In this work, we make the curated data available as AVID and introduce new measures and a benchmark to study the effects of target size, sensor misalignment, and image degradation.

AVID contains 32,040 corresponding RGB-infrared frame pairs from 115 ground-based surveillance videos, with 35,952 annotated bird and drone instances. The images have a resolution of 320×256 pixels, and most targets are very small. About 68.1% of the annotations have a geometric-mean bounding-box side between 8 and 16 pixels, while 99.0% fall within the COCO small-object category. We provide the images, annotations, video identifiers, and code needed to prepare and evaluate the data.

To study the alignment between the two cameras, we introduce an object-relative alignment ratio and an alignment burden curve that measure displacement relative to target size. Among 300 sampled frame pairs, the median estimated shift is 1 pixel, but 32.3% have shifts greater than 5 pixels. For the 264 pairs with matching target annotations, 16.3% have a displacement larger than the median target side. We also introduce AVID-C, a perturbation benchmark with seven types of image degradation at five severity levels, which can be applied to RGB, infrared, or both. Using a fixed feature representation, we examine how target size, corruption, and misalignment affect the separation between birds and drones. For targets between 12 and 16 pixels, the fitted Bhattacharyya error bounds are 0.148 for RGB, 0.198 for infrared, and 0.046 for the combined features. These bounds describe class separation in the chosen representation rather than detector accuracy. By making AVID and AVID-C available, we provide data and evaluation tools for studying how small target size, weak alignment, and image degradation affect RGB-infrared drone surveillance.

1 Introduction

Detecting drones from ground-based surveillance cameras is difficult when they are far from the camera and occupy only a few pixels in an image. At this scale, birds and drones can look very similar, with little shape or texture to distinguish them. Even small amounts of noise or blur can hide the details needed for detection and classification. Previous studies have explored bird-drone classification using deep learning [1], as well as drone localization and tracking in surveillance videos [2, 3]. While these methods address important aspects of drone surveillance, their performance also depends on the quality of the images and the amount of information available about each target.

Using RGB and infrared cameras together can help address some of these challenges. RGB images capture visible appearance, texture, and background details, while infrared images provide thermal information that may be useful when visible contrast is poor. Recent work has also explored combining these modalities for drone localization [4]. However, using both cameras introduces additional challenges. Since the sensors have separate optical systems, the same target may appear at different positions in the two images. A displacement of five pixels may have little effect on a large object, but it can represent half the width of a drone that is only ten pixels wide. Noise, blur, uneven illumination, and camera movement can further affect the two modalities differently. These factors make it important to understand not only how well a model performs, but also how target size, sensor alignment, and image quality influence the information available for detection.

*Dharini Raghavan and Vishnu Rao contributed equally to this work.

The imagery used in this work was originally published by Svanström and colleagues as part of the Drone Detection Dataset, which contains visible and infrared surveillance videos of drones and other flying objects [5, 6]. We used this dataset to curate bird-drone surveillance imagery for the IEEE ICIP 2025 Video and Image Processing Cup [7]. The curated data provided corresponding RGB and infrared views of distant aerial targets. However, the original challenge did not specifically examine how target size, spatial differences between the sensors, or image degradation affect these observations. Studying these properties is particularly important for small targets, where even minor changes in image quality or alignment can remove a significant amount of useful information.

In this paper, we build on our earlier VIP Cup curation by making the bird-drone data available as AVID and introducing additional tools to study these challenges. We prepare the dataset using the Train and Validation splits from the challenge, where the correspondence between RGB and infrared frames can be established from the available records. We remove duplicate noise-augmented images while preserving the original video identifiers, frame order, and annotations. The resulting dataset contains 32,040 RGB-infrared frame pairs from 115 surveillance videos, with 35,952 annotated instances, including 15,471 birds and 20,481 drones. Along with the dataset, we introduce new measures of sensor alignment, a controlled perturbation benchmark, and an analysis of how these factors affect the separation between birds and drones.

We first study the size of the targets and the alignment between the RGB and infrared images. In the native 320×256 frames, 24,489 annotations, or 68.1%, have a geometric-mean bounding-box side between 8 and 16 pixels. Nearly all targets, 99.0%, fall within the COCO small-object category. This makes it difficult to distinguish between different target sizes using the standard COCO categories alone. We therefore examine the objects at finer scales and measure sensor displacement relative to target size. To do this, we introduce an object-relative alignment ratio and an alignment burden curve that describe how often the estimated shift exceeds a given fraction of the target size. Among 300 sampled RGB-infrared pairs, 32.3% have an estimated displacement greater than 5 pixels. For the 264 pairs with matching target annotations, 16.3% have a displacement larger than the median target side. These results show why alignment should be considered relative to the size of the observed object rather than measured only in pixels. The alignment measures and estimation procedure are described in Section 4.2.

To study the effects of image degradation, we introduce AVID-C, a perturbation benchmark designed for the small targets in AVID. It includes seven perturbation families at five severity levels: Gaussian noise, speckle noise, salt-and-pepper noise, Gaussian blur, uneven illumination, motion blur, and camera instability. Each perturbation can be applied to RGB alone, infrared alone, or both modalities. This allows us to examine how degradation in one sensor affects the available information while the other remains unchanged. We also provide the code and parameters used to generate these perturbations so that the same experiments can be reproduced without storing separate copies of the corrupted images.

We further study how target size, image degradation, and sensor misalignment affect the distinction between birds and drones. For this analysis, we extract patches around the annotated targets and represent them using a fixed set of low-dimensional features. We then estimate the overlap between the bird and drone feature distributions for RGB, infrared, and their combined representation. By repeating this analysis across different target sizes, perturbations, and controlled shifts, we examine when combining the two modalities provides additional information and how that information changes as image conditions worsen. These measurements describe class separation within the chosen feature representation and density models rather than the accuracy of a trained detector.

Our goal is to provide a dataset and benchmark that allow researchers to study these effects separately. By reporting target size, sensor alignment, and the type and severity of image degradation, results on AVID can be examined in greater detail than with a single overall performance score.

The main contributions of this paper are:

1. **AVID, a curated RGB-infrared dataset for tiny aerial targets.** We make the bird-drone surveillance data curated for the IEEE ICIP 2025 VIP Cup available as AVID, containing 32,040 corresponding RGB-infrared frame pairs from 115 videos and 35,952 existing target annotations. We provide the images, annotations, video identifiers, and code used to prepare the dataset.
2. **New measures for studying weak RGB-infrared alignment.** We introduce an object-relative alignment ratio, an alignment burden curve, and a summary score to measure sensor displacement relative to target size. These measures help describe the effect of spatial misalignment when the observed objects occupy only a few pixels.
3. **AVID-C, a benchmark for studying image degradation.** We introduce seven perturbation families at five severity levels, with the option to apply them independently to RGB, infrared, or both modalities. We provide the code and parameters needed to reproduce these perturbations and evaluate their effects.
4. **An analysis of target scale, sensor degradation, and alignment.** We examine how bird-drone class separation changes with target size, image perturbations, and cross-modal displacement. Using a fixed feature representation, we compare RGB, infrared, and combined features and study how the results depend on feature dimension.

Table 1: Comparison of AVID with related visible-infrared and UAV datasets. The pairing column describes the correspondence provided between modalities. Paired or synchronized observations do not necessarily imply pixel-level alignment.

Dataset	Main target/view	Modalities	Pairing	Main characteristic
KAIST [8]	pedestrian, ground	RGB + thermal	synchronized	multispectral pedestrian detection
LLVIP [9]	pedestrian, ground	visible + IR	paired	low-light pedestrian imagery
M3FD [10]	multi-class, ground	visible + IR	paired	image fusion and object detection
VTUAV [14]	generic target, aerial	visible + thermal	paired video	visible-thermal object tracking
Anti-UAV300 [16]	UAV, surveillance	RGB + IR	both streams released	long-range UAV surveillance
MMFW-UAV [17]	fixed-wing UAV, air-to-air	visible + thermal	one-to-one	multi-sensor and multi-view observations
MM-UAV [18]	UAV tracking	RGB + IR + event	synchronized	tracking with three sensor modalities
UAVNet-MS [19]	small UAV	RGB + MSI	synchronized	multispectral UAV recognition
UAV-CB [20]	UAV, low-altitude	RGB + thermal	paired	UAV detection in complex backgrounds
AVID (ours)	bird/drone, surveillance	RGB + thermal	weakly aligned pairs	tiny targets, measured alignment, and controlled perturbations

2 Related datasets and benchmark design

2.1 Visible-infrared and UAV datasets

Several datasets have been developed to study the use of visible and infrared images in computer vision. The KAIST multispectral pedestrian dataset provides synchronized visible and thermal images for pedestrian detection [8]. LLVIP contains paired visible and infrared images captured under low-light conditions [9], while M3FD includes scenes from different environments for image fusion and object detection [10]. The FLIR ADAS dataset provides visible and thermal images for object detection in road scenes [11]. Although these datasets support research on combining information from different sensors, they mainly focus on pedestrians, vehicles, and other objects in ground-level scenes. Detecting distant birds and drones presents a different challenge because the targets often occupy only a small number of pixels.

Several datasets also focus on UAV detection and tracking, although their camera setups and objectives differ. VisDrone provides images and videos captured from cameras mounted on drones for detecting and tracking objects on the ground [12]. In contrast, ground-based drone surveillance requires detecting the drone itself, often against the sky or a complex background. DroneRGBT uses RGB and thermal imagery for aerial crowd counting [13], while VTUAV provides visible and thermal video sequences for object tracking [14]. The Anti-UAV benchmark focuses more directly on detecting and tracking drones from surveillance cameras and includes visible and infrared observations [15, 16].

Previous studies have also developed methods for drone recognition and tracking in surveillance videos. Selvi et al. studied the classification of birds and drones using deep learning [1]. SetNet was introduced for drone localization and zero-shot tracking [2], while IR-Setnet addressed localization and tracking in distorted infrared surveillance videos [3]. ModalFuseNet further explored the use of sensor fusion for drone localization [4]. These studies focus on developing methods to detect, classify, or track aerial targets. In this work, we focus on the data used for such tasks and study how target size, sensor alignment, and image degradation affect the information available in RGB and infrared observations.

More recent datasets have explored other combinations of sensors and UAV observation conditions. MMFW-UAV contains multi-sensor and multi-view observations of fixed-wing UAVs [17]. MM-UAV combines RGB, infrared, and event-camera data for UAV tracking [18], while UAVNet-MS introduces RGB and multispectral imagery for recognizing small UAVs [19]. UAV-CB focuses on detecting UAVs against complex backgrounds using RGB and thermal images [20]. These datasets provide different settings for studying UAV perception. Our work focuses specifically on small birds and drones observed from ground-based cameras, where the RGB and infrared frames correspond to the same scene but may have spatial differences.

The imagery used in AVID originates from the Drone Detection Dataset published by Svanström et al., which provides visible and infrared surveillance videos of drones and other flying objects [5]. We previously used this imagery to curate bird-drone surveillance data for the IEEE ICIP 2025 Video and Image Processing Cup. In the present work, we make the curated data available as AVID, with corresponding RGB-infrared frames, existing target annotations, and preserved video information. We also study properties that were not examined separately in the original challenge. In particular, we measure target size and the spatial displacement between the two camera views, and introduce AVID-C to study how different types of image degradation affect each modality. The aim is to provide both the data and the tools needed to examine these factors when developing and evaluating drone detection methods.

Table 1 compares AVID with other visible-infrared and UAV datasets, focusing on their sensing modalities, observation settings, and main objectives.

2.2 Corruption and data-centric benchmarks

Image quality can change because of sensor noise, blur, lighting conditions, or camera movement. These changes may affect the performance of detection and tracking methods, even when the scene and target remain the same. To study such effects, Hendrycks and Dietterich introduced ImageNet-C, which evaluates recognition models under different image corruptions and severity levels [21]. Similar benchmarks have since been developed for other tasks. UAV-C evaluates how common corruptions affect aerial object tracking [22], while MapBench studies the effect of sensor corruption on multi-sensor map construction [23]. These benchmarks show the value of testing image degradation under controlled conditions rather than relying only on performance measured using clean images.

We follow this approach in AVID-C, but design the perturbations for the small targets and paired RGB-infrared observations in AVID. The benchmark contains seven perturbation families at five severity levels, covering noise, blur, uneven illumination, and camera instability. We choose the perturbation strengths for the native 320×256 images because a small amount of blur or displacement can remove much of the visible detail from a target that is only a few pixels wide. We also allow each perturbation to be applied to RGB alone, infrared alone, or both. This makes it possible to study how the two modalities respond to image degradation separately and whether the unaffected sensor retains useful information.

Our work is also related to research on dataset quality and evaluation. DataPerf introduced benchmarks for studying data-centric machine learning [24], while DataComp examined how dataset construction and selection influence model performance [25]. WhyShift studied the importance of identifying the sources of distribution shift when evaluating machine learning models [26]. Mayo et al. examined dataset difficulty by relating visual recognition performance to the time needed to interpret images [27]. These studies motivate examining the characteristics of the data alongside the performance of models trained on them.

In this work, we study three properties of the curated RGB-infrared data: target size, alignment between the two camera views, and the effect of image degradation. We measure each property separately and examine how it changes the overlap between bird and drone feature distributions. These measurements complement conventional detection and classification metrics by providing information about the image conditions under which a model is evaluated. Researchers can therefore use AVID and AVID-C to compare methods across target sizes, alignment conditions, and perturbation types, rather than relying only on a single overall score.

3 Dataset construction and release scope

3.1 Source and scope

The images used in this work come from the Drone Detection Dataset developed by Svanström, Alonso-Fernandez, and Englund, which contains visible and infrared surveillance videos of drones and other flying objects [5, 6]. We previously curated the bird-drone surveillance data from this collection for the IEEE ICIP 2025 Video and Image Processing Cup, which focused on drone detection, tracking, and payload identification [7]. In this paper, we build on that curation by making the bird-drone data available as AVID and introducing additional methods to study sensor alignment and image degradation. Our work focuses on the bird-drone surveillance data and does not include the separate payload-identification task.

To prepare AVID, we use the Train and Validation splits from the VIP Cup, where the correspondence between RGB and infrared frames can be established from the available records. We refer to these frames as paired because they provide corresponding observations of the same scene. However, the images are not necessarily aligned at the pixel level, as the two cameras may observe the target at slightly different positions. We preserve the original video identifiers and frame indices to maintain the connection between frames and their source sequences.

We exclude the VIP Cup Test split because the RGB and infrared files use different numbering schemes, and the available records do not provide a reliable way to match the frames. Rather than introduce uncertain pairs, we restrict the dataset to the Train and Validation splits for which correspondence can be established.

3.2 Curation and annotations

The data prepared for the VIP Cup include clean frames and noise-augmented versions of the same observations. For AVID, we group the images using their frame identifiers and retain only the clean versions. This removes duplicate observations and provides a consistent starting point for the perturbation experiments introduced in this work. We then use AVID-C to apply controlled image degradation at different severity levels.

For each retained frame, we preserve its video identifier and frame index. This allows the original sequence order to be reconstructed and helps researchers separate videos when preparing training and evaluation sets.

We retain the bounding-box annotations and bird-drone class labels from the VIP Cup data. These annotations were part of the earlier curation and were not newly created for this study. The available source records do not fully describe the annotation

Table 2: Composition of the AVID dataset. The counts include the Train and Validation splits curated from the VIP Cup data after removing duplicate noise-augmented frames. RGB-infrared correspondence is established from the available frame records. Instance counts refer to the existing bird and drone bounding-box annotations.

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

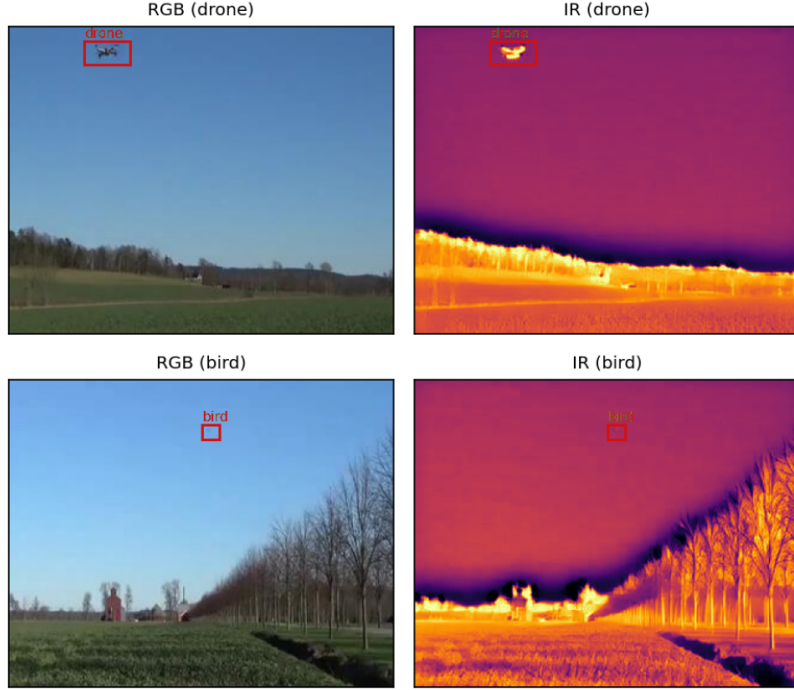


Figure 1: Representative RGB-infrared frame pairs from AVID. The top row shows a drone, and the bottom row shows a bird. Red bounding boxes indicate the annotated targets. The RGB and infrared images provide corresponding observations of the same scene, but the target’s appearance and position may differ between the two modalities.

and review procedures, so we use the existing labels without making additional assumptions about how they were produced.

3.3 Dataset composition

The resulting AVID contains 32,040 corresponding RGB-infrared frame pairs from 115 surveillance videos. As shown in Table 2, the Train split contains 28,790 pairs from 102 videos, while the Validation split contains 3,250 pairs from 13 videos.

Of the 115 videos, 76 contain drones and 39 contain birds. The sequences range from 50 to 329 paired frames, with a median length of 306 frames and a mean of 278.6 frames. The dataset includes 35,952 annotated target instances, consisting of 15,471 birds and 20,481 drones.

Figure 1 shows representative RGB and infrared frames containing a drone and a bird. In both examples, the targets occupy only a small part of the image, and their appearance differs between the two modalities. The same target may also appear at slightly different positions because the cameras are not perfectly aligned. These differences are important when working with small aerial targets, where even a few pixels of displacement can affect how information from the two sensors is combined.

4 The geometry of the dataset

4.1 A concentrated tiny-object regime

One of the main challenges in AVID is the small size of the birds and drones. Most targets occupy only a few pixels, leaving very little information about their shape or appearance. To understand how small these objects are, we first examine the distribution of their bounding-box sizes.

Using the COCO object-size convention [28], 35,581 of the 35,952 annotated instances, or 99.0%, fall within the “small” category. The remaining 371 instances are classified as “medium,” and none are “large.” Since nearly all objects belong to the same category, the standard COCO grouping does not capture the differences in target size within this dataset.

We therefore use the geometric-mean side length of each bounding box to describe target size more precisely:

$$s = \sqrt{w_{\text{px}} h_{\text{px}}}, \quad (1)$$

where w_{px} and h_{px} are the bounding-box width and height in pixels. This gives the side length of a square with the same area as the original bounding box, allowing us to represent each target using a single size measurement.

Figure 2a shows the distribution of target sizes in AVID. The largest group falls between 12 and 16 pixels, containing 17,706 annotated instances. When the 8 to 12 pixel range is also included, 24,489 instances, or 68.1% of the dataset, have a geometric-mean side between 8 and 16 pixels. Only 103 instances are smaller than 8 pixels, and no annotated target has a geometric-mean side larger than 48 pixels.

This concentration of small objects is important when studying the relationship between RGB and infrared images. A displacement of only a few pixels can be substantial when the target itself is around 10 or 15 pixels wide. For this reason, we examine not only the displacement between the two camera views, but also how large that displacement is relative to the size of the target.

4.2 Weak RGB-infrared alignment

Although the RGB and infrared frames in AVID correspond to the same scene, the two views are not necessarily aligned at the pixel level. The cameras use separate optical systems, so an object may appear at slightly different positions in the corresponding images. This difference becomes particularly important for small targets, where even a modest displacement can affect the information available when combining the two modalities.

To estimate the spatial difference between the RGB and infrared frames, we use normalized mutual information (NMI). We first convert each RGB frame to grayscale and then search for the translation that maximizes its NMI with the corresponding infrared frame. The search considers integer translations between -12 and 12 pixels along both the horizontal and vertical axes.

A displacement measured in pixels alone does not indicate how significant it is for the observed target. For example, a five-pixel shift represents half the width of a target that is 10 pixels wide, but only one tenth of a target that is 50 pixels wide. We therefore introduce an object-relative alignment ratio that expresses the estimated displacement in terms of target size.

For frame pair i , let d_i denote the magnitude of the estimated global translation between the RGB and infrared images, 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

$$\rho_i = \frac{d_i}{s_i}. \quad (2)$$

This ratio describes how large the estimated displacement is relative to the target. For example, $\rho_i = 0.5$ means that the displacement is half the median target side, while $\rho_i > 1$ means that it exceeds the median target side.

To examine how frequently different levels of misalignment occur, we introduce the alignment burden curve:

$$\hat{A}(\tau) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}[\rho_i > \tau]. \quad (3)$$

Here, n is the number of frame pairs with matching target annotations, τ is a chosen alignment threshold, and $\mathbf{1}[\cdot]$ is an indicator function. The curve gives the fraction of frame pairs whose estimated displacement exceeds τ times the median target side. For example, $\hat{A}(0.5)$ represents the fraction of pairs with a displacement greater than half the target size.

We also summarize the burden curve using

$$\hat{B}_T = \frac{1}{T} \int_0^T \hat{A}(\tau) d\tau = \frac{1}{nT} \sum_{i=1}^n \min(\rho_i, T). \quad (4)$$

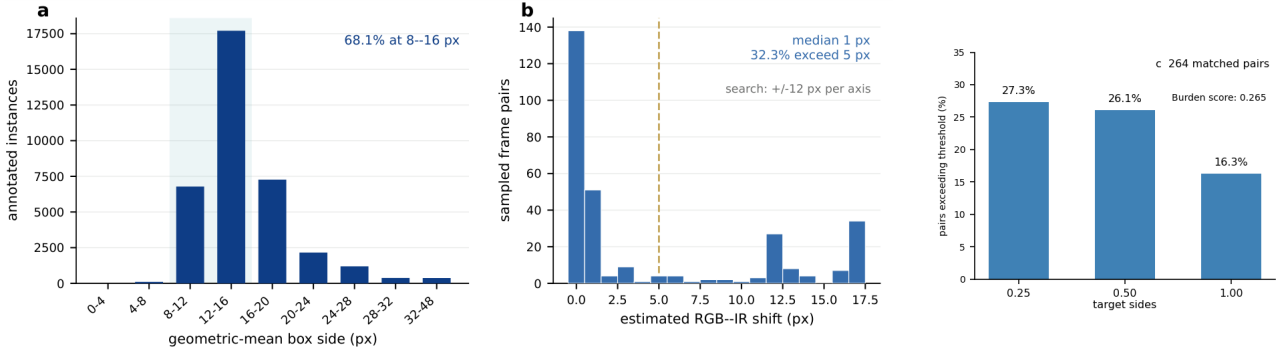


Figure 2: Target size and RGB-infrared alignment in AVID. **a**, Distribution of geometric-mean target side lengths across the Train and Validation data. The highlighted 8–16 pixel range contains 68.1% of the annotated instances. **b**, Estimated RGB-infrared translation magnitude for 300 sampled frame pairs. The search is limited to ± 12 pixels along each axis, corresponding to a maximum measurable translation magnitude of 16.97 pixels. **c**, Alignment burden for the 264 pairs with matching target annotations. The fractions exceeding one quarter, one half, and one full median target side are 27.3%, 26.1%, and 16.3%, respectively. The unit-scale alignment burden is $\hat{B}_1 = 0.265$.

This measure summarizes the distribution of object-relative displacements up to a chosen threshold T . In this work, we use $T = 1$, so ratios greater than one are clipped at one before averaging. The resulting score provides a single measure of alignment burden while the full curve shows how the displacements are distributed.

We estimate the RGB-infrared displacement for a sample of 300 frame pairs. 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. About 46.0% of the sampled pairs have an estimated integer translation of zero. However, 36.0% have shifts greater than 2 pixels, and 32.3% exceed 5 pixels, as shown in Figure 2b. These results indicate that although many corresponding frames have little estimated displacement, a substantial fraction show larger differences between the two camera views.

The estimated displacement is also limited by the range of the translation search. Since we consider shifts of up to 12 pixels along each axis, the largest translation magnitude that can be represented is

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

This means that a pair assigned a shift near the search boundary may have a larger displacement than the method can estimate. The maximum measured value of 16.97 pixels therefore reflects the search limit rather than the largest possible misalignment between the cameras.

For 264 of the 300 sampled frame pairs, matching target annotations are available to calculate the object-relative alignment ratio. The median target side in this subset is 14.18 pixels. The mean alignment ratio is 0.297, while the median is $\rho = 0$, reflecting the large number of pairs with zero estimated integer translation.

The burden curve provides a more detailed picture of these displacements. Among the 264 annotation-matched pairs, 27.3% have an alignment ratio greater than 0.25, meaning that the estimated displacement exceeds one quarter of the median target side. Similarly, 26.1% exceed half the target side, and 16.3% have a displacement larger than the full median target side. Thus, for approximately one in six of these pairs, the estimated shift between RGB and infrared is larger than the target itself when measured using its median geometric-mean side length.

Figure 2c shows the alignment burden at these three thresholds. Using Eq. 4, the unit-scale alignment burden is

$$\hat{B}_1 = 0.265. \quad (6)$$

This value summarizes the object-relative displacement after ratios above one have been clipped. It provides a compact measure of the alignment burden, while the individual threshold values show how frequently larger shifts occur.

These measurements show why alignment is an important consideration when working with tiny aerial targets. A displacement that appears small in the full image can be large relative to the object being detected. Expressing alignment in terms of target size therefore provides a more meaningful description of the spatial differences between the two modalities.

The estimated shifts describe global translations between the RGB and infrared frames. They do not provide exact pixel correspondences for individual targets or account for more complex differences such as parallax, lens distortion, or spatially varying displacement. Nevertheless, the object-relative alignment ratio and burden curve offer a practical way to describe the weak alignment present in the sampled frames of AVID.

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

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°)

5 AVID-C: a controlled perturbation benchmark

5.1 Perturbation families and severity levels

Apart from sensor misalignment, image quality also plays an important role in detecting small birds and drones. Noise can make a target difficult to distinguish from the background, while blur can remove details about its shape. Changes in illumination and camera movement can further affect how clearly a target appears. These problems are particularly important when the object occupies only a few pixels, and their effects may differ between RGB and infrared images.

To study these effects, we introduce AVID-C, a benchmark that simulates common forms of image degradation in the RGB and infrared frames of AVID. We consider seven types of perturbations: Gaussian noise, speckle noise, salt-and-pepper noise, Gaussian blur, uneven illumination, motion blur, and camera instability. Each perturbation is applied at five severity levels to examine how increasing degradation affects the images. The perturbations are generated from the original, unmodified frames using predefined parameters, allowing the same conditions to be reproduced across experiments without storing separate versions of the degraded images.

We choose the severity levels based on the native 320×256 image resolution and the small size of the targets in AVID. Although ImageNet-C introduced a widely used approach for evaluating robustness under common image corruptions [21], its settings are not necessarily suitable for very small objects. For example, a level of blur that has little effect on a large object may remove most of the visible details of a drone that is only 10 pixels wide. We therefore select perturbation strengths with the target sizes in AVID in mind. Table 3 lists the parameters used at each severity level.

For each perturbation, we consider three settings: degrading RGB alone, degrading infrared alone, and degrading both modalities. This allows us to examine how the two sensors respond when image quality changes. In particular, applying a perturbation to one modality while keeping the other unchanged helps us study whether the unaffected sensor retains useful information when the other becomes less reliable.

Since different perturbations may affect RGB and infrared differently, we recommend reporting results separately for each perturbation type, severity level, and affected modality. Reporting only an overall average could hide important differences in how the two sensors respond to image degradation.

5.2 Relating the perturbation scale to clean imagery

To understand how the simulated perturbations compare with the original images, we first examine the noise already present in the clean RGB and infrared frames. This is particularly useful for Gaussian noise, where the first severity level should introduce a change beyond the estimated noise in the original images.

We estimate the noise level using 200 frames from each modality and the no-reference method proposed by Immerkaer [29]. The method uses the Laplacian response of an image to estimate noise without requiring a noise-free reference. However, image texture and other fine details can influence the result. We therefore treat these values as approximate estimates of image noise rather than direct measurements of the physical sensor noise.

The median full-frame noise estimates are 0.302 intensity units for RGB and 0.638 for infrared on the 0–255 intensity scale. In comparison, the first Gaussian-noise severity level in AVID-C uses a standard deviation of 5 intensity units. This is approximately 16.6 times the estimated noise level in RGB and 7.8 times that in infrared.

We also repeat the estimation using only the upper third of each image, where relatively uniform sky regions are common. The median noise estimates in these regions are 0.027 intensity units for RGB and 0.047 for infrared, both lower than the corresponding full-frame estimates.

Figure 5i compares the estimated noise levels in the clean images with the first Gaussian-noise severity level. The simulated

noise is higher than the estimated noise in both modalities, indicating that even the lowest severity level introduces additional degradation relative to the original frames. Although these estimates do not provide a physical calibration of the cameras, they help establish how the chosen perturbation levels relate to the noise already present in the dataset.

6 A fixed-representation analysis of the dataset

6.1 Measuring class structure under controlled variation

We next examine how target size, image degradation, and sensor misalignment affect the differences between birds and drones. Rather than evaluating a particular detection model, we study the visual information available in the annotated target regions. This allows us to examine how the two classes differ under changing image conditions without introducing differences caused by model architecture or training.

For each annotated target, we extract a square image patch centered on its bounding box, with $1.5\times$ padding around the target. We convert the patches to grayscale, resize them to 24×24 pixels, and flatten them into feature vectors. We then apply principal component analysis (PCA) separately to the RGB and infrared patches, using components learned from the clean images. We retain eight principal components for each modality, giving an 8-dimensional representation for RGB and another for infrared. To study the two modalities together, we concatenate their features into a 16-dimensional representation.

The PCA components are learned once from the clean images and remain unchanged throughout the experiments. This is important because we want to measure how the target features change under corruption or misalignment, rather than compare representations learned separately for each condition. Figure 3 illustrates the analysis, from extracting target patches to comparing the resulting feature distributions.

6.2 Class overlap in the fixed representation

To measure how well birds and drones can be distinguished in the chosen feature space, we compare the distributions of their extracted features. When the two classes have similar feature distributions, distinguishing them becomes more difficult. When their distributions are more separated, the features provide stronger evidence for distinguishing one class from the other.

Let Z represent the extracted features and let $Y \in \{0, 1\}$ denote the bird or drone class. We model the features of each class using a Gaussian distribution:

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

Here, μ_y and Σ_y are the estimated mean and covariance matrix for class y .

We use the Bhattacharyya distance to measure the separation between the two fitted Gaussian distributions:

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

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

The Bhattacharyya distance can also be used to obtain an upper bound on the classification error. For class priors π_0 and π_1 , this bound is

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

We assume equal priors for birds and drones and denote the resulting fitted error bound by $\hat{B}$. A smaller value indicates greater separation between the two classes under the fitted Gaussian model. Importantly, this value describes the features extracted from known target locations. It is not the measured accuracy or error rate of a trained detection system.

Since the result may depend on the Gaussian assumption, we also calculate an optimized Chernoff bound and a kernel-density estimate (KDE). The Chernoff method searches for the parameter that gives the tightest bound between the fitted class distributions, while KDE provides a comparison without assuming Gaussian class densities. We use these three approaches to examine class separation for targets in the 12–16 pixel range, which contains the largest number of annotated instances in AVID.

6.3 Comparing single and paired modalities

One of the main reasons for using RGB and infrared cameras together is that the two sensors may capture different information about the same target. We therefore compare the class separation obtained from RGB alone, infrared alone, and their combined features.

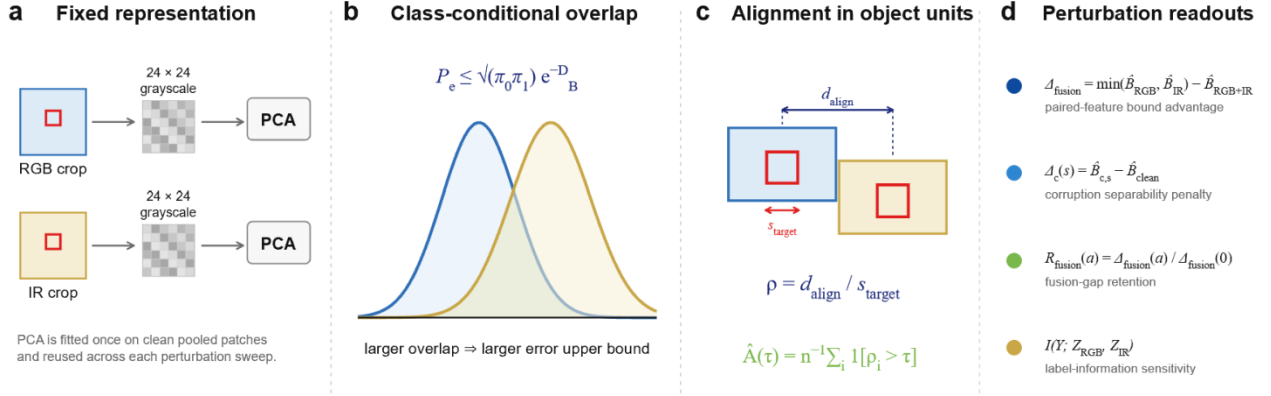


Figure 3: Overview of the fixed-representation analysis used to study AVID. **a**, Annotated RGB and infrared target patches are converted to 24×24 grayscale images and projected into PCA spaces learned from the clean data. **b**, Bird and drone feature distributions are compared using Bhattacharyya and Chernoff error bounds, together with a kernel-density estimate as a non-Gaussian comparison. **c**, Cross-modal displacement is measured relative to target size using $\rho = d_{\text{align}} / s_{\text{target}}$. **d**, The resulting measurements are used to compare RGB, infrared, and combined features and to examine the effects of image degradation and increasing sensor misalignment.

To measure the difference, we define the fusion gap as

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

Here, $\hat{B}_{\text{RGB}}$ and $\hat{B}_{\text{IR}}$ are the fitted Bhattacharyya error bounds for the individual modalities, while $\hat{B}_{\text{RGB+IR}}$ is the bound obtained from their combined features.

A positive fusion gap means that the combined representation has a lower fitted error bound than either individual modality. A value close to zero indicates little difference between the combined representation and the better-performing individual modality. A negative value means that the combined features produce a higher bound than the better individual modality.

This comparison must also account for the number of features used. In our main analysis, RGB and infrared each have eight PCA components, while their combined representation has 16. The additional dimensions may influence the fitted error bound, independently of the information contributed by the second sensor. We therefore interpret the fusion gap within the chosen representation and examine the effect of feature dimension separately in Section 7.

6.4 Effect of corruption

We next use the same feature representation to examine how simulated image degradation changes the separation between birds and drones. For each perturbation in AVID-C, we compare the fitted Bhattacharyya error bound with the corresponding value obtained from clean images.

For corruption type c at severity level s , we define the corruption penalty as

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

Here, $\hat{B}_{c,s}$ is the fitted error bound after applying the perturbation, and $\hat{B}_{\text{clean}}$ is the bound for the same evaluation setting without added degradation.

A positive value indicates that the fitted error bound increases after corruption, suggesting that the bird and drone features have become harder to separate. A negative value indicates a lower bound after corruption. This can occur because some image transformations, such as mild smoothing, may remove variation that does not help distinguish the two classes.

We calculate the corruption penalty separately for RGB and infrared using the clean baseline for each modality. This allows us to compare how different perturbations affect the information available from the two sensors.

6.5 Sensitivity to cross-modal shift

We also examine how spatial misalignment affects the separation obtained from the combined RGB-infrared features. Since the cameras are not perfectly aligned, the same target may appear at different positions in the two images. To study this effect under controlled conditions, we introduce additional shifts to the infrared target patches while keeping the RGB patches unchanged.

Let a denote the maximum horizontal and vertical displacement allowed in the experiment. We measure how much of the original fusion gap remains as the shift range increases. For this purpose, we define fusion-gap retention as

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

Here, $\Delta_{\text{fusion}}(a)$ is the fusion gap measured with an added shift budget of a , while $\Delta_{\text{fusion}}(0)$ is the gap without additional displacement.

A retention value of 1 means that the fusion gap is unchanged from the reference condition. Values between 0 and 1 indicate that some of the initial gap has been lost as the shift increases. This allows us to examine how sensitive the combined representation is to additional spatial differences between the two modalities.

Together, these measurements allow us to study different aspects of the dataset. The Bhattacharyya and Chernoff bounds describe class separation in the fitted feature distributions. The fusion gap compares the combined RGB-infrared representation with the better individual modality. The corruption penalty measures how class separation changes under simulated image degradation, while fusion-gap retention shows how the combined representation responds to increasing misalignment. By keeping the feature representation fixed, we can examine these effects under consistent experimental settings.

7 Dataset behavior under scale, corruption, and alignment

7.1 Target scale shapes class separation

We first examine how target size affects the separation between birds and drones in the fixed feature representation. Figure 4a shows the fitted Bhattacharyya error bounds for RGB, infrared, and their combined features across different target sizes. A lower bound indicates that the two classes are more easily separated within the chosen representation.

Across the size ranges with sufficient examples from both classes, combining RGB and infrared features generally produces lower fitted bounds than using either modality alone. This difference is particularly clear for targets between 12 and 16 pixels, which make up the largest group in AVID. At this scale, the Bhattacharyya bounds are $\hat{B}_{\text{RGB}} = 0.148$ for RGB, $\hat{B}_{\text{IR}} = 0.198$ for infrared, and $\hat{B}_{\text{RGB+IR}} = 0.046$ for the combined representation. The corresponding fusion gap is 0.103, as shown in Figure 4b. These results suggest that the two modalities provide more distinguishable bird and drone features when considered together, although the magnitude of the difference depends on the feature representation.

We also examine whether this result holds when class overlap is estimated using other methods. Figure 4c compares the Bhattacharyya bounds with optimized Chernoff bounds and kernel-density estimates (KDE) for the 12–16 pixel range. The Chernoff bounds are 0.131 for RGB, 0.194 for infrared, and 0.040 for the combined features. The corresponding KDE estimates are 0.165, 0.177, and 0.047. All three methods give the lowest estimated overlap for the combined representation at this target size. This agreement suggests that the observed ordering is not limited to the Gaussian Bhattacharyya calculation alone.

An important consideration is that the combined representation contains more features than either individual modality. In the main experiment, RGB and infrared each use eight PCA components, while their combined representation contains 16. As shown in Figure 4d, the fitted error bounds generally decrease as more components are included. We therefore examine whether the lower bound obtained from the combined features is partly explained by this increase in dimensionality.

To make the comparison more balanced, we construct an 8-dimensional combined representation using four PCA components from RGB and four from infrared. For targets between 12 and 16 pixels, this representation gives a fitted bound of 0.140, compared with 0.148 for RGB alone and 0.198 for infrared alone, both using eight components. The combined features still produce the lowest bound, but the difference is considerably smaller than in the 16-dimensional comparison. This ordering also does not hold across every larger target-size range, indicating that the benefit of combining the modalities varies with target scale and the features used.

We perform an additional check by duplicating each single-modality feature vector to create a redundant 16-dimensional representation. Since the duplicated components contain no new information, this experiment helps reveal whether the fitted bounds are sensitive to the number of dimensions. After duplication, the RGB bound changes from 0.148 to 0.127, while the infrared bound changes from 0.198 to 0.191. These changes suggest that covariance estimation or regularization influences the fitted bounds, even when no additional information is introduced. Both values remain above the combined bound of 0.046, but the results show that the fusion gap should be interpreted within the chosen feature representation rather than as a direct measure of improved detection performance.

7.2 RGB and infrared respond differently to corruption

We next examine how simulated image degradation affects the separation between birds and drones. Using AVID-C, we apply seven types of perturbations at five severity levels and compare their effects on RGB and infrared features. The experiments

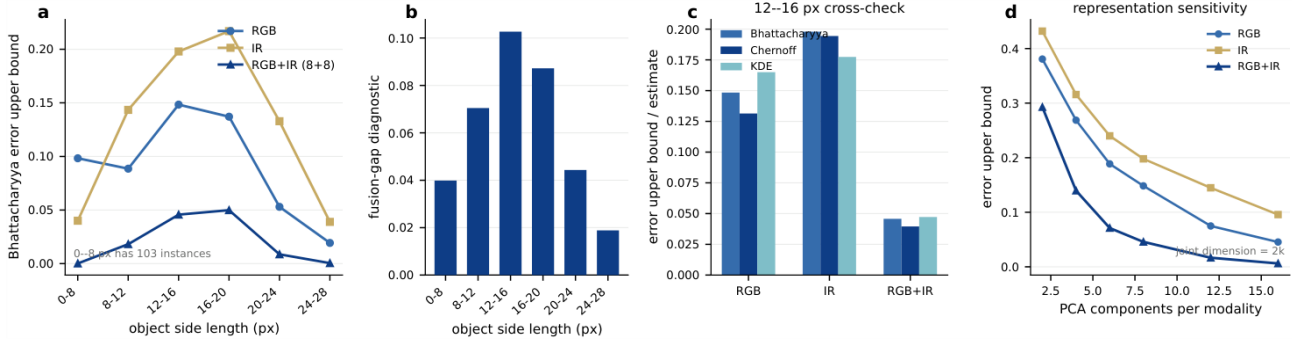


Figure 4: Effect of target size on bird-drone class separation. **a**, Fitted Bhattacharyya error bounds across target-size intervals for RGB, infrared, and the combined representation using eight PCA components from each modality. The 0–8 pixel interval contains only 103 annotated instances and should be interpreted with caution. **b**, Fusion gap calculated using Eq. 11 for different target sizes. **c**, Comparison of Bhattacharyya bounds, optimized Chernoff bounds, and KDE estimates for targets between 12 and 16 pixels. **d**, Effect of the number of PCA components on the fitted error bounds. The combined representation uses k components from each modality, giving a total of $2k$ features.

use the same 1,500 annotated target instances across all conditions, with the PCA representation kept unchanged. Figure 5a–g shows how the fitted Bhattacharyya bounds change from the clean images through severity level 5.

For this 1,500-instance subset, the clean-image bounds are 0.148 for RGB and 0.162 for infrared. These values differ from those reported for the 12–16 pixel target-size subset because the two experiments use different groups of examples.

As the perturbation strength increases, the fitted bounds generally rise, although the changes differ across perturbation types and between the two modalities. At severity level 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 illumination, 0.222 for motion blur, and 0.224 for camera instability. The corresponding infrared bounds are 0.198, 0.222, 0.189, 0.204, 0.245, 0.215, and 0.224.

The largest differences between RGB and infrared occur under the three noise perturbations. Gaussian noise, speckle noise, and salt-and-pepper noise produce much larger increases in the RGB bound than in the infrared bound. For example, at severity level 5, speckle noise increases the RGB bound from 0.148 to 0.375, while the infrared bound increases from 0.162 to 0.222. This suggests that, within the chosen feature representation, the infrared features retain greater bird-drone separation under these simulated noise conditions.

The differences are smaller for blur and camera-related perturbations. At severity level 5, Gaussian blur gives bounds of 0.186 for RGB and 0.204 for infrared, while motion blur produces values of 0.222 and 0.215. Camera instability results in a bound of approximately 0.224 for both modalities. These results show that neither sensor responds in the same way to every type of degradation.

When averaged across all seven perturbation families at severity level 5, the fitted error bound increases by 0.132 for RGB and 0.052 for infrared relative to their respective clean baselines (Figure 5h). The mean severity-5 bound is approximately 1.90 times the clean RGB value and 1.32 times the clean infrared value. Although these averages indicate a larger overall change in the RGB features for the perturbations considered here, the individual results are more informative because the differences depend strongly on the type of degradation.

Not all perturbations produce a steady increase in the fitted error bound. Under Gaussian blur, for example, the RGB bound initially decreases before increasing at higher severity levels. One possible explanation is that mild smoothing removes high-frequency variation that does not help distinguish birds from drones, while stronger blur begins to remove useful target details. Although the fitted bounds do not establish the exact cause of this behavior, the result shows that increasing the strength of an image perturbation does not necessarily produce a corresponding increase in estimated class overlap at every level.

7.3 Cross-modal shift weakens paired separation

We also examine how additional spatial misalignment affects the combined RGB-infrared representation. In this experiment, we keep the RGB target patches unchanged and introduce controlled shifts to the corresponding infrared patches. For each shift budget, horizontal and vertical offsets are sampled within the specified per-axis range. We repeat each nonzero shift condition five times while keeping the feature representation fixed.

Table 4 reports the results for shift budgets up to 12 pixels along each axis. We use this range to match the translation search described in Section 4.2. The shift budget represents the maximum allowed displacement along each axis rather than a fixed displacement applied to every target.

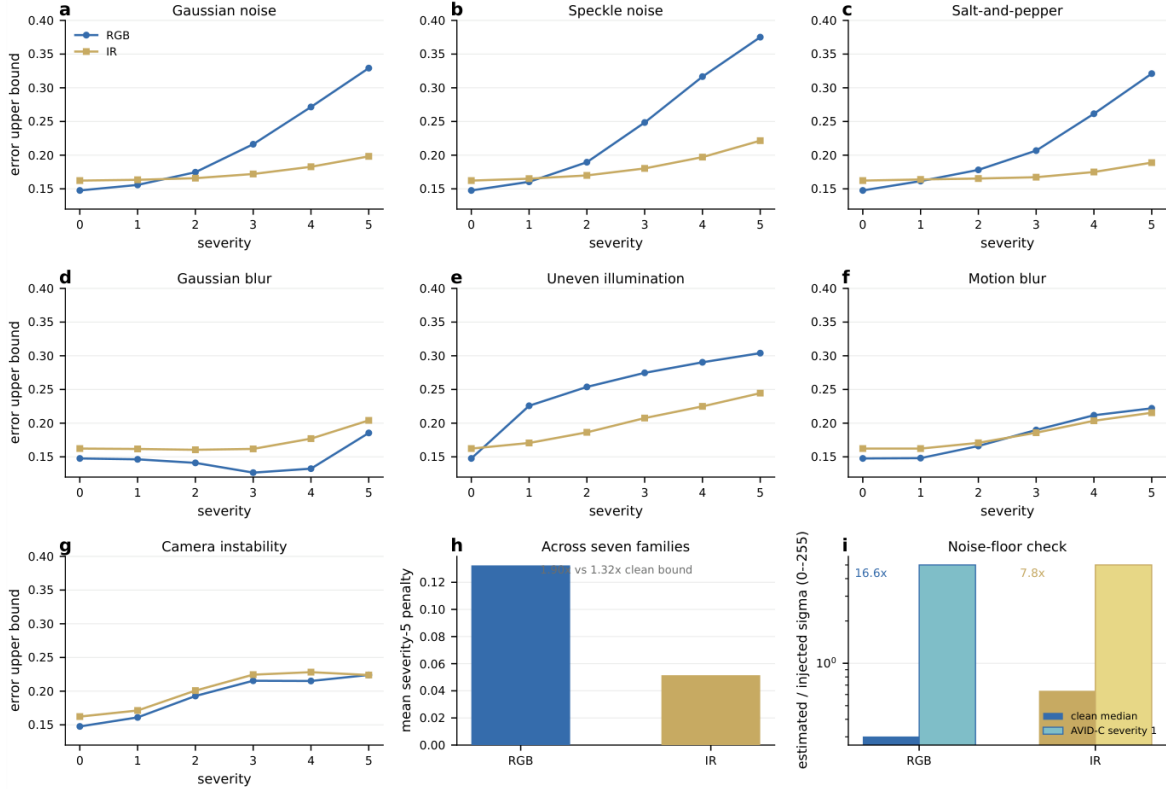


Figure 5: Effect of simulated image degradation on RGB and infrared class separation. Panels a–g show the fitted Bhattacharyya error bounds from clean images (severity 0) through severity level 5 for each of the seven perturbation families. h, Mean increase in the fitted error bound at severity level 5 relative to the corresponding clean baseline. i, Estimated noise levels in clean RGB and infrared images compared with Gaussian-noise severity level 1. The added noise has a standard deviation approximately 16.6 times the median RGB estimate and 7.8 times the median infrared estimate.

Without any additional shift, the fitted fusion gap is 0.111. As the allowed displacement increases, the gap gradually decreases, reaching 0.093 at the 12-pixel shift budget. The corresponding fusion-gap retention is

$$R_{\text{fusion}}(12) = 0.844, \quad (14)$$

indicating that approximately 84.4% of the initial fitted fusion gap remains at the largest shift budget considered.

The paired-feature bound also increases from 0.0370 to 0.0542 over the same range. Together, these changes indicate that additional displacement reduces the separation obtained from combining RGB and infrared features in this experiment. However, the result depends on the chosen feature representation and should not be interpreted as a direct prediction of detector accuracy.

It is also important to distinguish this experiment from the alignment measurements reported earlier. Figure 2c describes the estimated spatial differences already present in sampled RGB-infrared frames. Table 4, in contrast, shows how the fitted class separation changes when additional shifts are introduced under controlled conditions. The first measures a property of the dataset, while the second examines its effect on the chosen representation.

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

The experiments in this section examine three factors that can make bird-drone discrimination difficult. Target size determines how much visual detail is available to describe the object. Image degradation can obscure that detail through noise, blur, or other distortions. Spatial misalignment affects how the observations from RGB and infrared cameras can be combined.

These factors are particularly relevant to AVID because most annotated targets have a geometric-mean side between 8 and 16 pixels. At this scale, even a small amount of blur or a displacement of a few pixels can affect a substantial part of the target. The results also show that RGB and infrared do not respond identically to every perturbation, and that the separation obtained from combining their features can change as misalignment increases.

Taken together, these findings support evaluating small aerial targets under clearly described image conditions. Alongside

Table 4: Effect of additional infrared shift on fitted class separation. The shift budget a gives the maximum allowed displacement along each image axis, rather than a fixed radial displacement. Each nonzero condition is repeated over five trials.

Shift budget a (px)	Paired bound	Fusion gap	Gap retention
0	0.0370	0.1107	1.000
1	0.0383	0.1094	0.988
2	0.0414	0.1063	0.960
5	0.0472	0.1005	0.907
8	0.0507	0.0970	0.876
12	0.0542	0.0935	0.844

conventional detection or classification metrics, we recommend reporting target size, the affected modality, perturbation type and severity, and the estimated alignment between the sensors. This would help distinguish whether changes in performance are associated with limited target detail, image degradation, or differences between the two camera views.

8 Benchmark use and reporting

To make results on AVID easier to reproduce and compare, we recommend reporting the conditions under which each experiment is performed. In particular, target size, sensor alignment, and image degradation can influence performance in different ways. Reporting these details alongside conventional detection or classification metrics helps explain why a method performs well under some conditions but struggles under others.

8.1 Clean paired evaluation

When evaluating models on the original images without added perturbations, researchers should specify whether they use RGB, infrared, or both modalities. Any resizing or cropping should also be reported, since changing the native 320×256 resolution affects the apparent size of the targets and may influence detection performance.

The evaluation setup should identify the dataset split, number of frames, and video sequences used. The Train and Validation splits contain different source videos, and frames from the same video should remain within a single split to avoid overlap between training and evaluation. If new splits are created, we recommend separating the data by video rather than by individual frames.

The current version of AVID includes corresponding RGB-infrared frames from the VIP Cup Train and Validation splits. We do not include the original Test split because reliable frame correspondence could not be established from the available records. Therefore, the Test images should not be used as RGB-infrared pairs unless their correspondence is independently verified.

8.2 Scale-aware evaluation

Since most targets in AVID are very small, reporting performance only using the standard COCO object-size categories may hide differences between targets of different sizes. We recommend evaluating models across narrower size intervals using the geometric-mean side length of each bounding box:

$$s = \sqrt{w_{\text{px}} h_{\text{px}}}, \quad (15)$$

where w_{px} and h_{px} represent the bounding-box width and height in pixels.

For AVID, we use the intervals 0–8, 8–12, 12–16, 16–20, 20–24, 24–28, 28–32, and 32–48 pixels. Results should include the number of annotated instances in each interval together with the corresponding performance metric. This is especially important for the 0–8 pixel range, which contains only 103 instances and may not provide enough examples for reliable comparisons.

For tracking and other sequence-level tasks, target size can be summarized using the distribution of bounding-box sizes across frames or a sequence-level statistic such as the median target side. This helps distinguish sequences containing consistently small targets from those in which the target appears larger or changes size over time.

8.3 Corruption evaluation

When evaluating methods using AVID-C, we recommend reporting the type of simulated degradation, its severity level, and the modality to which it was applied. Each perturbation can be tested under three conditions:

Table 5: Recommended reporting details for experiments using AVID and AVID-C. Reporting the evaluated data, target sizes, sensor conditions, and experimental settings helps make results easier to reproduce and compare.

Item	What to report
Input	RGB, infrared, or both modalities; original or resized image resolution
Data subset	Dataset split, video identifiers when relevant, and number of frames
Target scale	Geometric-mean bounding-box size intervals and number of instances
Corruption	Perturbation type, severity level, and affected modality or modalities
Alignment	Estimation method, search range, and object-relative alignment thresholds or burden curve
Aggregation	Results across target sizes and perturbation types, alongside any overall average
Fixed representation	Crop definition, PCA dimensions, class priors, and density model for comparisons with Section 6

1. RGB is degraded while infrared remains unchanged.
2. Infrared is degraded while RGB remains unchanged.
3. Both RGB and infrared are degraded.

These settings allow researchers to examine how a method performs when image quality decreases in one sensor or both. When perturbations are applied to both modalities, the experiment should also specify whether the random changes are generated independently or share the same random realization.

We recommend presenting results for each perturbation type across all five severity levels rather than reporting only an average. As observed in our Gaussian-blur experiments, increasing the strength of a perturbation does not always produce a steady change in the fitted class-separation measure. Examining the full severity range can therefore reveal behavior that may be missed when results are combined into a single score.

8.4 Alignment-aware evaluation

Because the RGB and infrared frames are not necessarily aligned at the pixel level, we also recommend reporting the estimated displacement between the two views when an alignment measurement is available. A shift measured only in pixels may be difficult to interpret for very small targets. We therefore use the object-relative alignment ratio introduced in Section 4.2:

$$\rho = \frac{d_{\text{align}}}{s_{\text{target}}}, \quad (16)$$

where d_{align} is the estimated displacement between the modalities and s_{target} is the target size used for normalization.

To describe the extent of misalignment, we recommend using the thresholds $\tau \in \{0.25, 0.5, 1.0\}$. These indicate whether the estimated displacement exceeds one quarter, one half, or one full target side, respectively.

The alignment burden at these thresholds can be reported as

$$\hat{A}(0.25), \quad \hat{A}(0.5), \quad \hat{A}(1.0), \quad (17)$$

along with the unit-scale burden $\hat{B}_1$, which summarizes the object-relative displacement up to one target side.

The alignment estimation method and its search range should also be specified. In our analysis, we estimate global translation using normalized mutual information with a search range of ± 12 pixels along each axis. Other alignment methods may produce different displacement estimates, so their results should be interpreted using the corresponding estimation procedure.

Expressing displacement relative to target size makes it easier to understand how significant a shift is for small objects. However, comparisons between studies still require a consistent definition of target size and a clear description of how alignment was estimated.

8.5 Comparison with the fixed-representation analysis

The class-separation results reported in Section 6 depend on the features and statistical models used in the analysis. Researchers wishing to compare their results directly with these values should follow the same experimental setup, including the target-crop definition, 24×24 grayscale patches, PCA dimensions, equal class priors, and density-estimation methods.

Other feature representations may provide useful insights, but their fitted error bounds and fusion gaps should not be compared directly with ours without accounting for differences in the representation and estimation procedure.

Table 5 summarizes the information we recommend including when reporting results on AVID and AVID-C.

9 Data access and licensing

The original surveillance videos were made available through the Drone Detection Dataset under the CC0 1.0 Universal dedication [6]. This allows the source material covered by that dedication to be reused, modified, and redistributed. The original dataset authors also request that users cite their research when using the data [5].

We make the bird-drone data curated for the VIP Cup available as AVID at <https://huggingface.co/datasets/Dharini24/AVID>. The dataset contains corresponding RGB-infrared frames, existing bird and drone annotations, and video identifiers that preserve the connection to the original sequences. The AVID listing specifies CC BY-NC 4.0 for the released dataset package. This designation does not change the CC0 status of the underlying source imagery, which remains available under its original terms.

The scripts used to prepare the dataset, the alignment analysis, and the code for generating the simulated perturbations in AVID-C are available at https://github.com/rdharini2001/ICIP_2025_Data_Release. The code is released under the MIT license. We encourage researchers to cite both the original Drone Detection Dataset and this work when using AVID, and to report the dataset version used in their experiments.

10 Responsible use

AVID contains RGB and infrared surveillance images of birds and drones. It is intended to support research on small-object detection, sensor fusion, tracking, alignment, and robustness to image degradation. The dataset does not include labels identifying individuals or describing personal attributes.

Research on drone surveillance can support applications such as airspace monitoring and safety, but it also raises concerns about how surveillance technologies may be used. The ability to distinguish birds from drones in a benchmark does not establish that a system is suitable for deployment. Real-world use requires additional evaluation of detection errors, false alarms, operating conditions, legal requirements, and the consequences of incorrect decisions. Human oversight is also important when detection results may lead to actions affecting people or property.

We retain birds as a separate labeled class because they are a common source of confusion in drone surveillance. At long distances, birds and drones may have similar appearances, especially when only a few pixels describe the target. Treating birds only as background would make it harder to evaluate this distinction, which is an important part of the task.

11 Limitations

Although AVID and AVID-C provide a way to study small aerial targets under different image conditions, several limitations should be considered when interpreting the results.

The original Test split is not included in the paired dataset. We use the VIP Cup Train and Validation splits because their RGB and infrared frame correspondence can be established from the available records. The original Test files do not provide sufficient information to match the two modalities reliably. We therefore exclude them rather than introduce uncertain frame pairs into the dataset.

The original annotation process is not fully documented. The bird and drone bounding boxes come from the data curated for the VIP Cup. We retain these annotations rather than create new labels for this paper. However, the available records do not fully describe how the annotations were assigned, reviewed, or resolved when disagreements occurred. As a result, we cannot report annotator agreement or provide an independent measure of annotation uncertainty.

Alignment is measured on a subset of the dataset. We estimate RGB-infrared displacement using 300 sampled frame pairs, of which 264 have matching target annotations for the object-relative analysis. These measurements describe the sampled observations and should not be interpreted as alignment estimates for all 32,040 frame pairs.

The alignment analysis considers only global translation. We estimate one horizontal and vertical displacement for each RGB-infrared frame pair using normalized mutual information. This provides a measure of the overall spatial difference between the two images, but it cannot capture local variations caused by parallax, lens distortion, differences in viewing geometry, or target depth. The estimated shift therefore does not represent exact pixel correspondence between the modalities.

The alignment search has a limited range. The translation search is restricted to ± 12 pixels along each image axis, corresponding to a maximum measurable displacement magnitude of 16.97 pixels. When the estimated shift reaches this limit,

the actual displacement may be larger. Consequently, the upper end of the measured alignment distribution may be affected by the search boundary, and the results should not be interpreted as a complete calibration of the cameras.

Class-separation results depend on the feature representation. Our main analysis uses eight PCA components from each modality and Gaussian models fitted to the bird and drone feature distributions. The combined representation contains more dimensions than either individual modality, which can affect the fitted error bounds. When we compare representations with the same total number of dimensions, the difference between paired and single-sensor features becomes smaller. Duplicating single-sensor features also changes the fitted bounds despite adding no new information, indicating sensitivity to covariance estimation or regularization. Although the KDE analysis provides a comparison without the Gaussian assumption, it still depends on the chosen features. These results describe class separation within the stated representation and should not be interpreted as detector accuracy.

The analysis uses known target locations. We extract image patches around the existing ground-truth bounding boxes. This allows us to study the visual information available for distinguishing birds from drones, but it does not measure the difficulty of locating the targets in full surveillance frames. The analysis also does not evaluate temporal tracking, association between frames, or the performance of a complete detection system.

AVID-C represents a selected set of image degradations. The benchmark simulates seven common types of degradation involving noise, blur, illumination changes, motion, and camera instability. These perturbations allow us to study their effects under controlled conditions, but they do not cover every problem that can arise in real surveillance. Other factors, including atmospheric interference, thermal sensor artifacts, compression failures, lens contamination, rolling-shutter effects, and more complex camera motion, are not included in the current benchmark.

Increasing perturbation severity does not always increase difficulty. The severity levels in AVID-C describe the strength of the simulated image transformation, not a guaranteed increase in classification or detection difficulty. As observed in the Gaussian-blur experiment, mild degradation can sometimes reduce the fitted class-overlap bound before stronger degradation increases it. The relationship between perturbation strength and model performance therefore needs to be examined rather than assumed.

The dataset does not cover all surveillance conditions. The 115 videos include different backgrounds, target appearances, and observation conditions, but they cannot represent every environment in which drone surveillance may be used. The dataset does not cover all bird species, drone designs, camera systems, weather conditions, viewing distances, or geographic locations. Evaluation on additional datasets and under different operating conditions is needed before drawing conclusions about how well a method generalizes beyond AVID.

12 Conclusion

In this work, we build on the bird-drone surveillance data curated for the IEEE ICIP 2025 Video and Image Processing Cup from the original Drone Detection Dataset [5]. We make this curated data available as AVID, containing 32,040 corresponding RGB-infrared frame pairs from 115 videos and 35,952 existing bird and drone annotations. We also introduce measures for studying sensor displacement relative to target size and AVID-C, a benchmark that simulates seven types of image degradation at five severity levels.

Our analysis shows why target size, alignment, and image quality are important when working with small aerial objects. About 68.1% of the annotated targets have a geometric-mean bounding-box side between 8 and 16 pixels. At this scale, a displacement of only a few pixels can be large relative to the target. Among the 264 sampled frame pairs with matching annotations, 16.3% have an estimated RGB-infrared shift greater than the median target side. We also find that simulated image degradation affects RGB and infrared features differently, particularly under noise perturbations.

Using a fixed PCA representation, we further examine how these conditions affect the separation between birds and drones. The combined RGB-infrared features generally produce lower fitted class-overlap bounds than the individual modalities at the most common target size, although the difference depends on feature dimension. We also observe that the fitted advantage of combining the two modalities decreases as additional spatial shifts are introduced. These findings describe the behavior of the chosen feature representation rather than the accuracy of a trained detection system.

By making AVID and AVID-C available, we aim to support further research on detecting and distinguishing small birds and drones using RGB and infrared observations. The dataset, alignment measures, and simulated perturbations allow researchers to examine how performance changes with target size, sensor displacement, and image quality. Reporting these

conditions alongside conventional evaluation metrics can provide a clearer understanding of when and why a method succeeds or fails.

Acknowledgements

We thank Fredrik Svanström, Fernando Alonso-Fernandez, and Cristofer Englund for collecting and publicly sharing the original Drone Detection Dataset [5]. We used this dataset to curate the bird-drone surveillance imagery 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 also thank the wider VIP Cup team for their contributions to organizing and preparing the challenge. The present work builds on this earlier curation by preparing the AVID release and introducing the alignment measures, AVID-C benchmark, and accompanying analysis.

References

- [1] 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)*, 2022. doi: 10.1109/MysuruCon55714.2022.9972589. URL <https://ieeexplore.ieee.org/abstract/document/9972589>.
- [2] 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.
- [3] Dharini Raghavan, S Sethu Selvi, 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)*, pages 3600–3604, 2023. doi: 10.1109/ICIPC59416.2023.10328384.
- [4] 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.
- [5] Fredrik Svanström, Fernando Alonso-Fernandez, and Cristofer Englund. A dataset for multi-sensor drone detection. *Data in Brief*, 39:107521, 2021. doi: 10.1016/j.dib.2021.107521.
- [6] DroneDetectionThesis. Drone-detection-dataset: Infrared, visible and audio data for drone detection. <https://github.com/DroneDetectionThesis/Drone-detection-dataset>, 2020. Original source repository, CC0 1.0 Universal license; accessed October 2026.
- [7] 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 October 2026.
- [8] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung 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.
- [9] 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 3496–3504, 2021. doi: 10.1109/ICCVW54120.2021.00389.
- [10] 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)*, pages 5802–5811, 2022.
- [11] Teledyne FLIR. FLIR thermal dataset for algorithm training. <https://oem.flir.com/solutions/automotive/adas-dataset-form/>, 2018.
- [12] 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.

- [13] Tao Peng, Qing Li, and Pengfei Zhu. RGB-T crowd counting from drone: A benchmark and MMCCN network. In *Asian Conference on Computer Vision (ACCV)*, 2020.
- [14] 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)*, pages 8886–8895, 2022.
- [15] 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. doi: 10.1109/TMM.2021.3128047.
- [16] Anti-UAV Project. Anti-UAV official repository and dataset releases. <https://github.com/ZhaoJ9014/Anti-UAV>, 2026. Project repository; accessed October 2026.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] 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.
- [21] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] John Immerkær. Fast noise variance estimation. *Computer Vision and Image Understanding*, 64(2):300–302, 1996. doi: 10.1006/cviu.1996.0060.
- [30] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, 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 [30] and provides additional information about the origin, composition, preparation, and availability of AVID.

A.1 Motivation

Where does the imagery come from?

The original images come from the Drone Detection Dataset collected and released by Svanström and colleagues [5]. We previously curated the bird-drone surveillance imagery from this collection for the IEEE ICIP 2025 Video and Image Processing Cup [7]. The present work builds on that curation and does not involve collecting new surveillance videos.

What is new in AVID and AVID-C?

We make the curated bird-drone data available as AVID, with corresponding RGB-infrared frame pairs, existing bounding-box annotations, and video identifiers. We also introduce measures for estimating sensor misalignment relative to target size and AVID-C, a benchmark that simulates common forms of image degradation under controlled conditions. The payload-identification task from the original VIP Cup is not included in this work.

A.2 Composition

What does the paired dataset contain?

AVID contains 32,040 corresponding RGB-infrared frame pairs from 115 surveillance videos, with 35,952 annotated bird and drone instances. The Train split includes 28,790 frame pairs from 102 videos, while the Validation split contains 3,250 pairs from 13 videos. The annotations consist of 15,471 bird instances and 20,481 drone instances.

The frames are paired based on their correspondence in the available records. This does not imply that the RGB and infrared images are perfectly aligned at the pixel level.

What is the image resolution?

Both RGB and infrared frames have a native resolution of 320×256 pixels. The dataset statistics and target-size measurements reported in this paper refer to this original resolution.

How long are the video sequences?

The 115 video sequences contain between 50 and 329 paired frames. The median sequence length is 306 frames, and the mean is 278.6 frames.

How small are the targets?

Most birds and drones occupy only a small region of the image. About 68.1% of the annotated targets have a geometric-mean bounding-box side between 8 and 16 pixels, while 99.0% fall within the COCO “small” object category. Table 6 provides a more detailed breakdown of the target sizes.

Table 6: Distribution of target sizes in AVID. Target size is measured using the geometric-mean side length of each annotated 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 dataset does not include the payload-identification data used in the VIP Cup. We also exclude the original Test split from the paired dataset because the available file identifiers do not establish reliable correspondence between its RGB and infrared frames.

A.3 Collection and annotation

What viewpoint does the dataset represent?

The images were captured using ground-based surveillance cameras observing birds and drones at a distance [5]. Unlike datasets collected from cameras mounted on drones, AVID focuses on distant aerial targets viewed from the ground.

Where do the bounding boxes come from?

We retain the bird and drone bounding-box annotations from the data curated for the VIP Cup. These labels were not newly created for this paper. The available records do not fully describe the original annotation and review process, so we do not report the number of annotators or an annotation agreement score.

Were multiple versions of the same frame present in the source data?

Yes. The VIP Cup Train and Validation data contained clean frames and noise-augmented versions associated with the same frame identifiers. When preparing AVID, we grouped these observations by frame identifier and retained the clean version. All dataset counts and statistics in this paper refer to these deduplicated frames. The simulated perturbations in AVID-C are generated separately from the clean images.

A.4 Access and maintenance

Where is the dataset maintained?

The dataset is available at <https://huggingface.co/datasets/Dharini24/AVID>. The project repository, including dataset preparation scripts, alignment analysis, and the implementation of AVID-C, is available at https://github.com/rdharini2001/ICIP_2025_Data_Release.

How should future corrections be handled?

Any corrections to frame pairing, annotations, or derived metadata should be recorded in a new dataset version so that earlier experimental results remain reproducible. If reliable RGB-infrared correspondence becomes available for the original Test split, those frames should be introduced in a separate version rather than changing the composition of the current paired dataset.

What license applies to the data and code?

The original source imagery was released under CC0 1.0 Universal [6]. The AVID listing specifies CC BY-NC 4.0 for the released dataset package, but this does not change the original CC0 terms that apply to the source imagery. The project code is released under the MIT license.

B Alignment measurement

This appendix provides additional details about how we estimate the spatial displacement between the RGB and infrared images. The analysis uses a sample of 300 corresponding frame pairs from the Train and Validation splits of AVID.

We first convert each RGB image to grayscale and resize the infrared image to the same spatial dimensions when necessary. We then estimate the global translation between the two images using normalized mutual information (NMI).

For each frame pair, we evaluate integer translations along the horizontal and vertical axes within the range

$$(d_x, d_y) \in [-12, 12]^2. \quad (18)$$

For each candidate translation, we compare the shifted images after excluding a 12-pixel border. This prevents pixels introduced at the image boundaries during translation from affecting the measurement.

We calculate NMI using a 32-bin joint intensity histogram:

$$\text{NMI}(A, B) = \frac{H(A) + H(B)}{H(A, B)}, \quad (19)$$

where $H(A)$ and $H(B)$ are the individual image entropies and $H(A, B)$ is their joint entropy.

We select the translation that produces the highest NMI value. The magnitude of the estimated displacement is then calculated as

$$d_{\text{align}} = \sqrt{d_x^2 + d_y^2}. \quad (20)$$

The 300 frame pairs are selected randomly from the paired Train and Validation data using a fixed selection. These frames provide a sample of the alignment conditions in AVID, but their distribution may not represent every frame in the dataset.

We also calculate a secondary displacement estimate using phase correlation on Sobel edge-magnitude images. This provides an additional comparison with the NMI estimates. However, neither method provides ground-truth registration. Their results may differ because infrared and RGB images have different intensity characteristics, and both approaches can be affected by regions with little texture.

To understand the significance of the estimated shift for a small target, we normalize the displacement by target size. For each sampled frame pair i , we use the median geometric-mean side length of its annotated targets, denoted by s_i . The object-relative alignment ratio is

$$\rho_i = \frac{d_i}{s_i}. \quad (21)$$

This ratio expresses the estimated displacement in units of the median target side. For example, $\rho_i = 0.5$ indicates that the estimated displacement is half the median target size.

We summarize the frequency of different relative displacements using the empirical alignment burden curve:

$$\hat{A}(\tau) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}[\rho_i > \tau]. \quad (22)$$

Here, n is the number of frame pairs with matching target annotations, and τ is the chosen target-relative threshold. The burden curve gives the fraction of pairs whose estimated displacement exceeds that threshold.

We also calculate a summary score for displacements up to one target side:

$$\hat{B}_1 = \int_0^1 \hat{A}(\tau) d\tau = \frac{1}{n} \sum_i \min(\rho_i, 1). \quad (23)$$

This score is the average object-relative displacement after ratios greater than one have been clipped at one. It provides a single summary value while the burden curve describes how the displacements are distributed.

Table 7 reports the alignment statistics for the sampled frame pairs. Of the 300 pairs, 264 have matching target annotations that allow the object-relative displacement to be calculated.

The estimated translation has a mean of 4.66 pixels, a median of 1.00 pixel, and a standard deviation of 6.44 pixels. Although 46.0% of the sampled pairs have an estimated integer translation of zero, 32.3% have shifts greater than 5 pixels.

For the 264 annotation-matched pairs, the median target side is 14.18 pixels. The alignment burden exceeds 0.25 target sides for 27.3% of these pairs, 0.5 target sides for 26.1%, and one full target side for 16.3%. The resulting unit-scale burden is $\hat{B}_1 = 0.265$.

Table 7: Summary of RGB-infrared alignment measurements. Global translation is estimated for 300 frame pairs, of which 264 have matching target annotations for the object-relative analysis. The maximum representable shift is determined by the translation search range and does not indicate the largest possible physical displacement between the cameras.

Quantity	Value
Number of measured shifts	300
Pairs with matched annotations	264
Pairs without matched annotations	36
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 (264 pairs)	14.18 px
$\hat{A}(0.25)$	0.273
$\hat{A}(0.50)$	0.261
$\hat{A}(1.00)$	0.163
$\hat{B}_1$	0.265

An important limitation of this analysis is the translation search range. Since shifts are restricted to ± 12 pixels along each axis, the largest displacement magnitude that can be represented is 16.97 pixels. A frame pair assigned a shift near this boundary may have a larger actual displacement.

These measurements should therefore be interpreted as estimates of global image translation within the chosen search range. They do not provide exact correspondence between individual targets or account for local spatial differences between the two camera views.

C Fixed-representation analysis details

C.1 Patch representation and scale analysis

The fixed-representation analysis uses the 35,952 annotated bird and drone instances in AVID. For each instance, we extract a square image region centered on the annotated bounding box, with $1.5\times$ padding around the target. We convert the extracted region to grayscale, resize it to 24×24 pixels, and flatten the pixel intensities into a feature vector.

We use principal component analysis (PCA) to reduce the feature dimension. Separate PCA bases are fitted to the clean RGB and infrared patches. In the main analysis, we retain eight components from each modality, giving an 8-dimensional RGB representation and an 8-dimensional infrared representation. We concatenate these two vectors to form the 16-dimensional paired representation.

All experiments use equal bird and drone class priors. The PCA components are learned from the clean images and remain fixed when we examine different target sizes, image perturbations, and additional spatial shifts.

To study the effect of target size, we group annotated instances according to their native geometric-mean bounding-box side length. We include a size interval in the class-separation analysis only when it contains at least 15 bird instances and 15 drone instances. This avoids estimating class distributions from intervals with too few examples from either class.

Table 8 reports the fitted Bhattacharyya error upper bounds for each supported target-size interval. Smaller bounds indicate greater separation between the fitted bird and drone feature distributions.

Table 8: Bhattacharyya error upper bounds across target sizes. The paired representation combines eight PCA components from RGB and eight from infrared. A dash indicates that the size interval does not meet the minimum sample requirement for both classes.

	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 smallest target-size interval, 0–8 pixels, contains only 103 instances, consisting of 52 birds and 51 drones. Although this interval meets the minimum sample requirement, its results should be interpreted more cautiously than those from intervals containing larger numbers of examples.

C.2 Chernoff and KDE comparison

The main analysis uses the Bhattacharyya error bound to compare the fitted bird and drone distributions. Since this measure depends on the Gaussian models used for the class features, we also calculate an optimized Chernoff bound and a kernel-density estimate (KDE) for comparison.

The Bhattacharyya bound corresponds to the $s = 0.5$ case of the Chernoff family. To obtain the optimized Chernoff bound, we evaluate 99 values of the Chernoff parameter s between 0.01 and 0.99 and retain the value that produces the smallest bound.

We also use KDE to estimate class overlap without assuming that the feature distributions are Gaussian. For the 12–16 pixel target-size interval, we draw 4,000 samples across the two classes and estimate the Bhattacharyya coefficient by importance sampling from an equal mixture of the fitted class densities.

We focus on the 12–16 pixel range because it contains the largest number of annotated instances in AVID. Table 9 compares the three estimates at this target size.

Table 9: Comparison of class-overlap estimates for targets between 12 and 16 pixels. Results are reported for the fitted Bhattacharyya bound, optimized Chernoff bound, and KDE-based estimate.

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

All three methods produce the lowest estimated class overlap for the combined RGB-infrared representation. This suggests that the ordering observed in the main analysis is not limited to the Gaussian Bhattacharyya calculation.

However, the estimates still depend on the chosen feature representation and the corresponding density-estimation procedures. They should therefore be interpreted as measures of class separation within this experimental setting rather than as detection error rates.

C.3 Representation-dimension controls

The main analysis combines eight RGB and eight infrared PCA components into a 16-dimensional representation. Because the individual modalities each use only eight components, the lower fitted bound obtained from the paired representation may partly reflect the larger feature dimension.

To examine this effect, we perform two additional comparisons.

First, we construct an 8-dimensional paired representation using four PCA components from RGB and four from infrared. This gives the paired and individual modalities the same total number of features, allowing a more direct comparison of the fitted bounds.

Second, we create redundant 16-dimensional representations by duplicating the eight-component RGB or infrared vectors. These duplicated features contain no additional information, but they allow us to examine whether the fitted bounds change simply because the feature dimension increases.

Table 10 reports the results for the 12–16 pixel target-size interval.

Table 10: Effect of feature dimension on fitted class separation at 12–16 pixels. The equal-dimension and duplicated-feature comparisons help identify how feature dimension influences the estimated bounds.

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

When all representations use eight dimensions, the combined RGB-infrared features give a fitted bound of 0.1396, compared with 0.1484 for RGB alone and 0.1979 for infrared alone. The paired representation still has the lowest bound, but the difference is much smaller than in the 16-dimensional comparison.

The duplicated-feature experiments also change the estimated bounds. Repeating the RGB features reduces the fitted bound from 0.1484 to 0.1272, while repeating the infrared features changes it from 0.1979 to 0.1913. Because duplication introduces no new information, these differences indicate sensitivity to the fitting procedure, including covariance estimation or regularization.

These checks show that the magnitude of the fusion gap depends on the selected feature dimension and statistical model. We therefore use the fitted bounds to compare conditions within a fixed representation rather than treating them as a general measure of the information gained by combining sensors.

C.4 Corruption analysis

To study the effects of simulated image degradation, we select 1,500 annotated target instances and use the same examples across all perturbation conditions. The PCA bases are fitted using clean patches and remain unchanged as the perturbation type and severity level vary.

We apply the seven perturbation families in AVID-C and compare the resulting Bhattacharyya error bounds with those obtained from the clean images. For this subset, the clean bounds are 0.1477 for RGB and 0.1623 for infrared.

Table 11 reports the fitted bounds at severity level 5 for each perturbation family.

Table 11: Bhattacharyya error upper bounds at severity level 5 of AVID-C. The corresponding clean-image bounds 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

The results show that different perturbations affect RGB and infrared features by different amounts. The largest differences between the modalities occur under Gaussian noise, speckle noise, and salt-and-pepper noise. In these cases, the fitted RGB bound increases more than the infrared bound.

The differences are smaller for Gaussian blur, motion blur, and camera instability. These observations show why it is useful to report the results for each perturbation family separately instead of relying only on an average across all types of

degradation.

As in the main analysis, these values describe the separation between fitted bird and drone feature distributions. They do not directly measure the performance of a trained detector under the simulated perturbations.

C.5 Controlled cross-modal shift

We also examine how additional spatial displacement affects the combined RGB-infrared representation. In this experiment, we keep the RGB target patch fixed and translate the corresponding infrared patch.

We define the shift budget a as the maximum allowed displacement along each image axis. The tested budgets are

$$a \in \{1, 2, 5, 8, 12, 17\} \text{ pixels.} \quad (24)$$

For each nonzero shift budget, we perform five trials. Within each trial, horizontal and vertical translations are drawn independently from the interval $[-a, a]$.

Because the extracted target patches are resized to 24×24 pixels, the sampled displacements are mapped from native-image coordinates into the resized patch space. We use a scale factor of $24/16 = 1.5$, corresponding to the dominant 12–16 pixel target-size range.

The main comparison includes shift budgets up to 12 pixels along each axis, matching the search range used in the measured alignment analysis. The 17-pixel condition is outside this range and is therefore not included in the main results table.

Table 12 shows how the fitted paired bound, fusion gap, and fusion-gap retention change as the shift budget increases.

Table 12: Effect of controlled infrared displacement on the paired feature representation. The shift budget specifies the maximum allowed displacement along each image axis rather than a fixed radial displacement. Nonzero conditions are evaluated over five trials.

Shift budget	Joint upper bound	Fusion gap	Gap retention
0	0.0370	0.1107	1.000
1	0.0383	0.1094	0.988
2	0.0414	0.1063	0.960
5	0.0472	0.1005	0.907
8	0.0507	0.0970	0.876
12	0.0542	0.0935	0.844

The paired Bhattacharyya bound increases from 0.0370 without additional displacement to 0.0542 at the 12-pixel shift budget. Over the same range, the fusion gap decreases from 0.1107 to 0.0935, leaving 84.4% of the initial gap.

This result indicates that increasing spatial displacement reduces the fitted class-separation advantage of the paired representation in the controlled experiment. However, the measurement is specific to the selected features and density models. It does not establish how a particular detection or tracking model would perform under the same shifts.

D Full-frame AVID-C examples

Figures 6 and 7 show examples of the seven simulated perturbation families applied to representative RGB and infrared frames.

For each perturbation, the examples show how the image changes as severity increases from level 1 to level 5. The images are presented at the native frame resolution so that the effect of each transformation can be viewed in relation to the small aerial target and its background.

These examples provide a visual reference for the perturbation parameters listed in Table 3. They also illustrate that the same type of simulated degradation may appear different in RGB and infrared images because the two modalities capture different scene information.

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

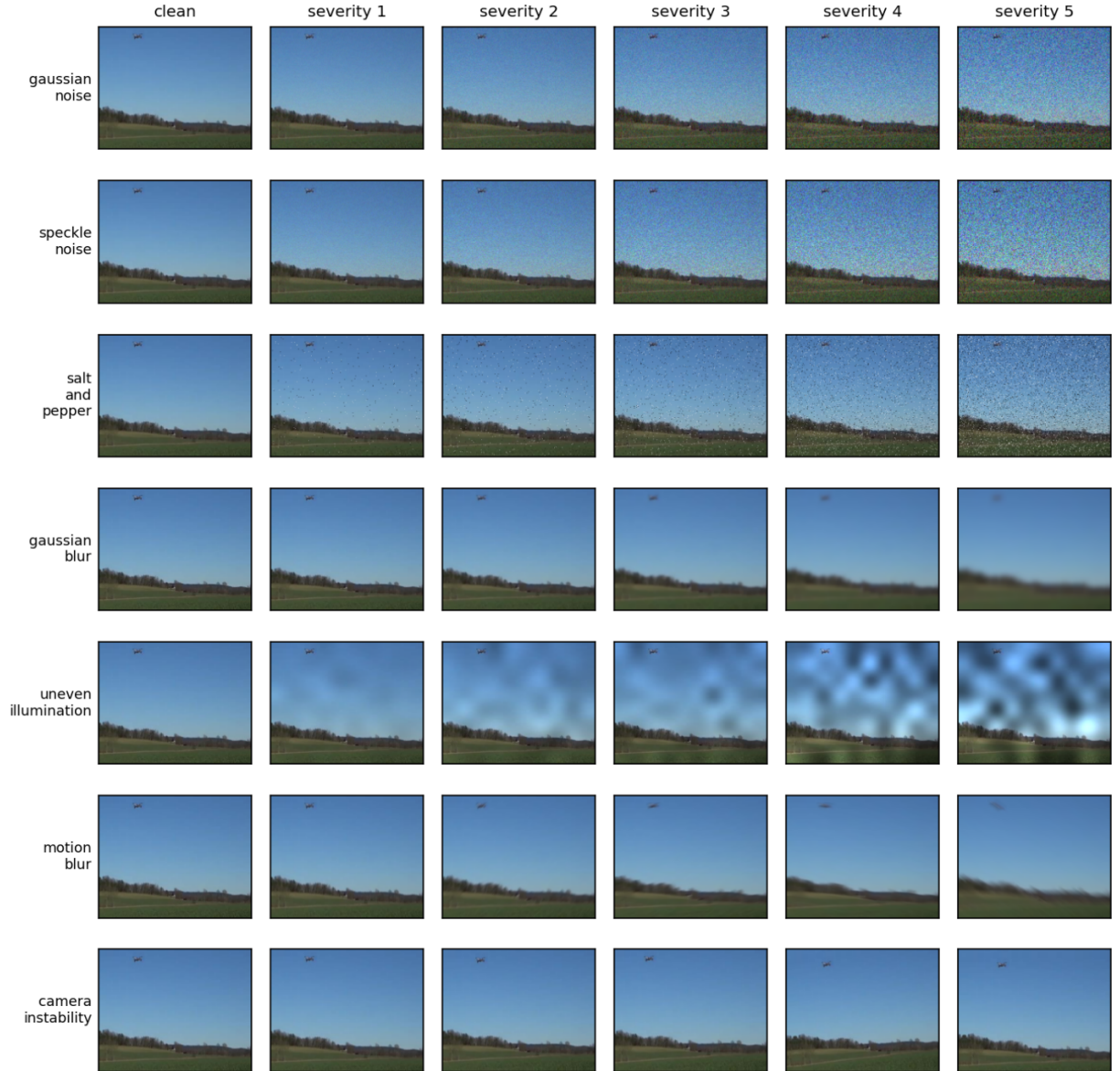


Figure 6: Examples of simulated image degradation in RGB frames. Each row shows one of the seven perturbation families in AVID-C, with severity increasing from left to right. The corresponding parameter values are listed in Table 3.

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

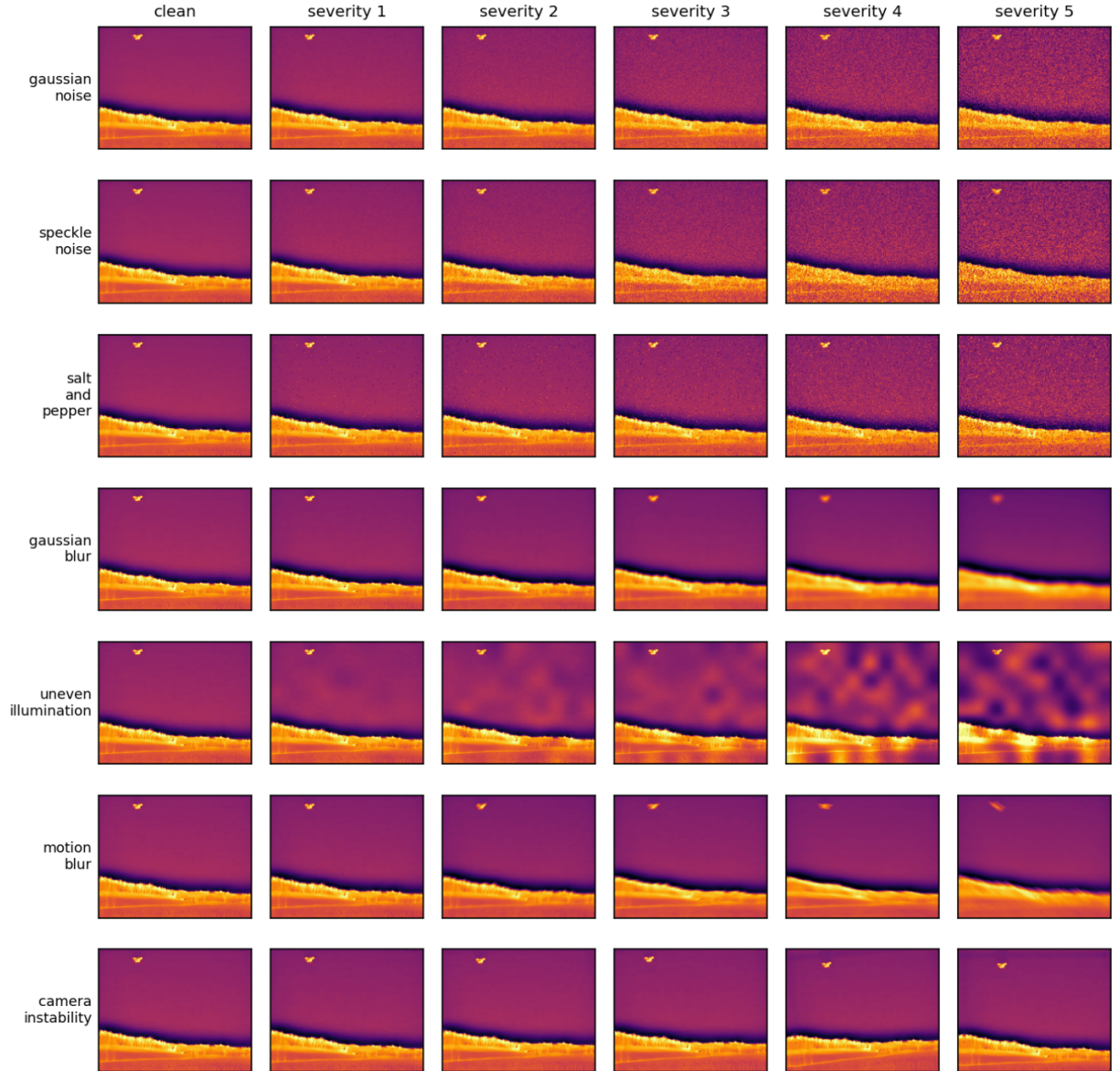


Figure 7: Examples of simulated image degradation in infrared frames. Each row shows one of the seven perturbation families across increasing severity levels, applied to the infrared imagery. The perturbation settings correspond to those used for the RGB examples.

E Additional dataset examples

Figure 8 shows additional RGB-infrared frame pairs selected from different source videos in AVID. The examples illustrate variation in scene backgrounds, target appearance, the number of visible targets, and thermal contrast.

Although all the images contain small aerial targets observed from ground-based surveillance cameras, their appearance can differ substantially across scenes. The surrounding background and the visibility of the target may also differ between RGB and infrared. These examples provide additional context for the dataset characteristics and the challenges discussed in the main paper.

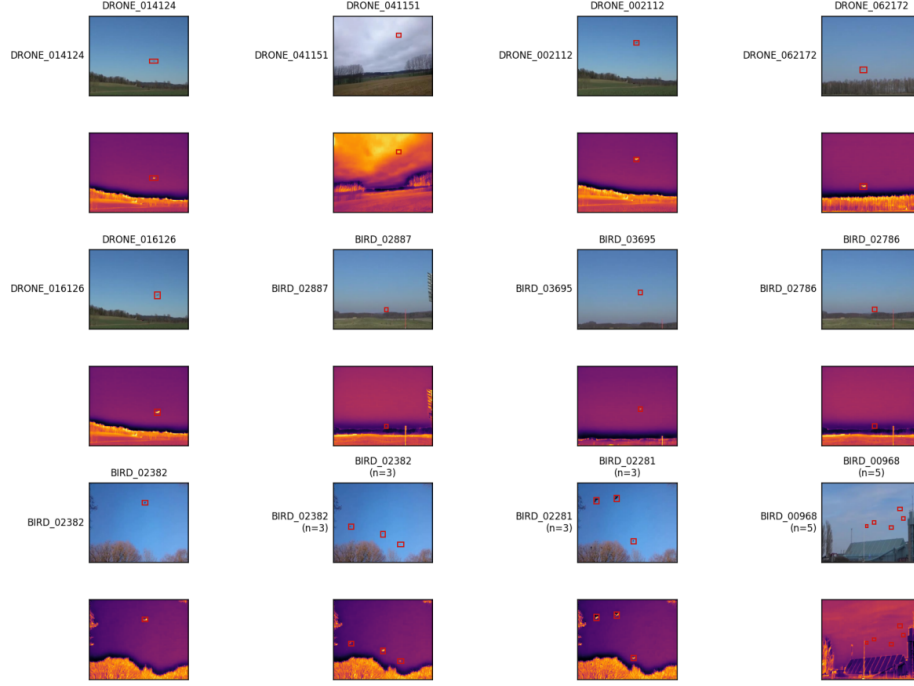


Figure 8: Additional RGB-infrared examples from AVID. The montage includes paired frames from different source videos, showing variation in background, target appearance, target count, and thermal contrast. Red bounding boxes indicate the annotated targets.