Meta-YOLO: Metadata-Guided Real-Time Object Detector in Aerial Imagery

Deukryeol Yoon, Seonghak Kim, Young Hwa Sung, Jinho Jung
Agency for Defense Development, Daejeon, Republic of Korea
{deukryeol, hakk35, yhsungadd, jinhojung}@korea.kr

Abstract

Aerial object detection is constrained by tiny targets, large scale variation, and strict real-time limits. It supports traffic monitoring, disaster response, and infrastructure inspection. Yet current detectors often ignore available platform metadata and process frames in isolation. This omission prevents receptive fields from adapting to scale variation and reduces accuracy. We propose META-YOLO *to exploit platform metadata for scale-aware aerial object detection in real time.* META-YOLO *injects normalized telemetry into spatial sampling to guide feature extraction. It modulates deformable convolution offsets using a spatial metadata map aligned with the image. This links visual features with platform state and enables receptive fields to adapt to object scale. Built on YOLOX,* META-YOLO *adds two modules: feature modulation and offset correction. Evaluated on 327K aerial frames with metadata,* META-YOLO *achieves up to +8.7 AP gains over YOLOX in lightweight regimes and consistently outperforms other recent detectors. It preserves real-time throughput with negligible overhead and improves accuracy without extra visual processing.*

1. Introduction

Aerial object detection aims to localize and classify objects from bird's-eye view imagery. This task is critical for various real-world applications, including traffic monitoring [30], disaster response [23], and environmental surveillance [24]. In particular, it serves as a foundational technology for enhancing the perceptual capabilities of aerial platforms such as unmanned aerial vehicles (UAVs), drones, and satellites.

Despite the rapid advances of general-purpose object detection in natural images driven by deep learning techniques [9, 25, 48], aerial object detection remains a challenging problem due to several unique characteristics of aerial imagery. First, objects often appear very small and exhibit large variations in scale. Second, densely distributed targets and cluttered backgrounds cause severe occlusion; thus hindering robust discrimination. Third, viewpoint variations induced by different altitudes and sensor orientations distort object appearances. Lastly, detectors on aerial platforms face limited computational capacity, which further compounds the difficulty of this task.

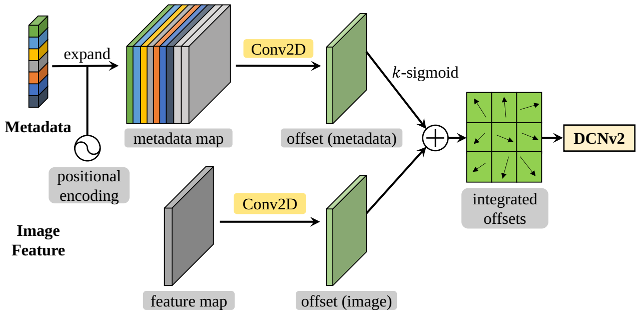

Figure 1. Architecture of the proposed metadata-guided offset modulation. Image features (bottom) and platform metadata (top) are individually processed by convolutional module, then fused to generate adaptive offsets for DCNv2. This design enables the convolutional kernel to adaptively align its receptive fields according to both image content and platform metadata.

Even with these challenges, aerial platforms inherently provide rich sensor and platform metadata that can serve as valuable auxiliary information for robust detection. For example, platform altitude, camera orientation, slant range, and field-of-view information are often recorded and standardized in formats such as Motion Imagery Standard Board (MISB) ST 0601 [1]. These metadata fields capture the physical context of image acquisition. They can be further leveraged to derive informative cues such as the ground sample distance (GSD), which relates image pixels to real-world resolution [8] and provides a prior for estimating object size. By integrating such information into the pipeline of lightweight detection models, we aim to enhance detection performance while preserving the computational efficiency required for real-time aerial platforms.

Previous attempts to exploit metadata have shown potential, but current approaches still face inherent limitations that restrict their effectiveness. Some works predict proxies for metadata from images to reduce scale confusion (*e.g.* GSD-based reasoning [16] and GSD feature embedding [43]). Other methods incorporate limited cues directly (*e.g.* altitude), using altitude-informed fusion [17] or adaptive normalization [21]. Recent work also fuses auxiliary information in small-target regimes through MLP-based conditioning [27]. However, most approaches leverage only a narrow set of metadata fields, and they do not align feature sampling with pixel-level scale or perspective. Distinct from prior approaches, we exploit the comprehensive standardized metadata defined in MISB ST 0601 and propose a spatially adaptive modulation mechanism that injects metadata into deformable layers for scale-aware feature extraction.

In this paper, we introduce META-YOLO, a real-time aerial object detector that injects metadata into spatial feature extraction through two core contributions. ① **(Metadata-Guided Modulation)** We normalize telemetry fields and construct a spatial metadata map aligned with image resolution. A modulation branch uses this map to compute additive offset corrections, guiding the learned sampling to reflect platform state such as altitude and orientation. (Figure 1) ② **(Scale-Aware Spatial Feature Modulation)** We modulate spatial sampling by learning offsets through deformable convolution, enabling dynamic receptive fields that adapt to object size and shape. This compensates for severe scale variation without distorting appearance, and aligns feature extraction with target geometry. By leveraging standardized MISB metadata, our design provides significant gains in lightweight settings, where limited capacity makes scale variation particularly problematic.

We conduct experiments demonstrating that META-YOLO consistently outperforms YOLO-series real-time detectors (YOLOX, YOLOv7, YOLOv8, and PP-YOLOE) across lightweight model sizes (Nano–Tiny and Small) while maintaining low computational cost. In particular, it achieves an 8.7% AP improvement in the Tiny regime, highlighting a strong efficiency–accuracy tradeoff for edge deployment. Beyond comparative benchmarks, we also conduct detailed analyses, including ablation studies on metadata utilization, comparisons of alternative modulation techniques, and evaluations of feature extractor designs, which together validate the effectiveness of our approach.

In summary, we make the following contributions:

- **System:** We introduce META-YOLO, a metadata-aware real-time detector that integrates platform context into spatial feature extraction.
- **Metadata Effectiveness:** We demonstrate that incorporating standardized Motion Imagery (MISB ST 0601) metadata into the detection pipeline significantly enhances the detection performance of lightweight models.
- **Modulation Method:** We design a metadata-guided modulation mechanism that leverages deformable convolution to adapt spatial receptive fields based on platform metadata, enabling scale-aware and context-aware feature extraction.
- **Experiments:** We achieve substantial accuracy gains in lightweight regimes with negligible computational cost, validating practicality for embedded aerial inference.

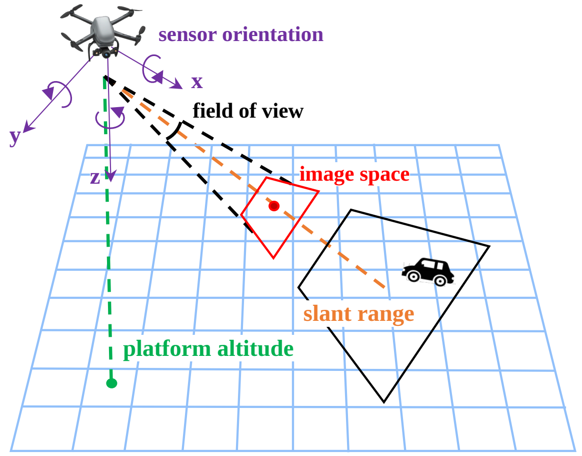

Figure 2. Platform metadata for aerial imagery following MISB ST 0601 specification. It includes platform altitude (green), sensor orientation angles (purple), horizontal and vertical fields of view, and slant range (orange), which together define the geometric relation between the platform, the onboard sensor, and the image.

2. Preliminaries

2.1. Platform Metadata for Aerial Imagery

Platform metadata accompanies aerial imagery captured by drones or UAVs. This auxiliary information contains sensor and platform specific parameters that cannot be inferred directly from the image content. Examples include viewing geometry, platform attitude, and sensor field of view. We leverage this metadata to guide spatial behavior in the detection pipeline.

We follow the MISB ST 0601 specification [1]. This widely adopted standard ensures consistent representation of motion imagery metadata, from which we extract the following attributes (Figure 2):

- **Slant range:** direct distance from the camera to the center point of the image.
- **Platform altitude:** height of the camera above the ground.
- **Sensor orientation:** sensor pose relative to the platform, including azimuth, elevation, and roll angles.
- **Field of View (FOV):** horizontal and vertical angular coverage of the sensor.

2.2. Deformable Convolution Network

Deformable Convolution Network v2 (DCNv2) [47] is an extension of standard convolution that attends to sparse spatial locations by learning both sampling positions and their importance. Unlike standard convolution, which samples input features on a fixed regular grid, DCNv2 introduces learnable sampling offsets and modulation scalars into the operation. The output at location p_0 is computed as follows:

$$y(p_0) = \sum_{p_n \in \mathcal{R}} \Delta m_n \cdot w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \quad (1)$$

Here, $\mathcal{R}$ denotes a regular grid (*e.g.* 3×3), Δp_n is the learnable offset for sampling location p_n, and $\Delta m_n \in [0, 1]$ is a modulation scalar predicted for each location. This formulation allows the network to adaptively learn both the sampling position and the contribution of each location. It achieves this by optimizing offsets and modulation scalars directly from the input features. In this work, we extend the range to $\Delta m_n \in [0, 2]$, so that the network can not only suppress but also amplify contributions from specific sampling locations (see Algorithm 1, line 6).

3. META-YOLO Model

META-YOLO is a real-time object detection model that integrates platform metadata to improve spatial feature representation. It enhances receptive fields by modulating the offset field of convolution modules within the feature pyramid network (FPN). We first outline the motivation behind this design. Then, we present the metadata-guided modulation module, which is the core component of our approach. Finally, we describe architectural modifications that enable scale-aware spatial feature extraction. The model is built on YOLOX [9], a widely used and practical object detector.

3.1. Motivation

META-YOLO enhances spatial feature representation by aligning receptive fields to object scale through metadata guided offset modulation. In aerial imagery, object scale varies across the image because of perspective distortion and slanted viewing angles. As shown in Figure 3, platform and sensor metadata (e.g., field of view, altitude, slant range) directly influence the object scale and contextual coverage of the scene. This provides the expected physical size of a target at each pixel location. We exploit this spatial prior to guide deformation of convolutional sampling patterns.

3.2. Metadata-Guided Modulation

Our metadata-guided modulation consists of three key steps: ① Normalization of raw metadata, ② Construction of a metadata map aligned with feature maps, and ③ Offset modulation that integrates metadata into deformable convolution. We describe each step in detail below.

Metadata Normalization To address distributional variance in metadata, we apply min–max normalization to each feature, mapping values to $[0, 1]$.

$$\tilde{m}_i = \frac{m_i - m_i^{\text{min}}}{m_i^{\text{max}} - m_i^{\text{min}}} \tag{2}$$

where m_i denotes the i-th raw metadata value, and m_i^{min}, m_i^{max} are the minimum and maximum of that feature. All values (*i.e.* minimum and maximum) are computed offline using the training dataset before model training.

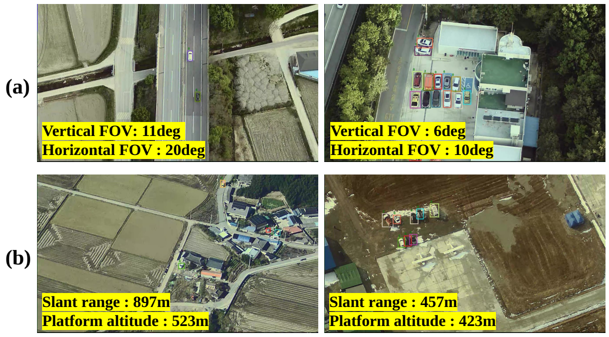

Figure 3. Examples of different metadata factors in aerial imagery. (a) Variation in sensor FOV (with similar platform altitude and slant range) changes the apparent size of objects and the contextual coverage of the scene. (b) Variation in platform altitude and slant range (with constant sensor FOV) alters object scale due to changes in ground sample distance. These examples highlight that platform and sensor metadata provide strong priors on object appearance and detection difficulty.

Metadata Map Construction We inject metadata into spatial feature extraction by constructing a metadata map aligned with the spatial dimensions of the intermediate feature maps. Each normalized metadata value is broadcast across spatial dimensions to form a constant map of size $H \times W$. We further append two positional encoding maps, corresponding to the x and y coordinates, in order to provide spatial context. This yields a metadata map $\mathbf{M} \in \mathbb{R}^{(N+2)\times H\times W}$, where N denotes the number of original metadata channels.

Metadata-Guided Offset Modulation To leverage platform metadata within convolutional modules, we modulate the sampling offsets of DCNv2 through a metadata-aware auxiliary branch (Figure 1). Specifically, we employ a lightweight convolutional module that takes the metadata map $\mathbf{M}$ as input and produces metadata-guided offset values Δp_n^{meta}. To ensure that the metadata contribution remains stable and bounded, we apply a scaled element-wise sigmoid activation $k \cdot \sigma(\cdot)$ to Δp_n^{meta}, constraining its values to the range $[0, k]$. In our implementation, the scaling factor is empirically set to $k = 2$ (see §4.3). We combine the metadata-aware modulation with the learnable image-based offset $\Delta p_n^{\text{image}}$, so that the final offset adapts to both image content and platform metadata:

$$\Delta p_n = \Delta p_n^{\text{image}} + k \cdot \sigma(\Delta p_n^{\text{meta}}) \tag{3}$$

where p_n refers to the learnable offset for sampling location described in §2.2.

Finally, the combined offset Δp_n in Equation 3 is substituted into the standard DCNv2 formulation in Equation 1. We refer to this extended formulation as *Metadata-Guided Deformable Convolution (MDCN)*, with the full procedure summarized in Algorithm 1.

Algorithm 1: Metadata-Guided Deformable Convolution (MDCN)

```
Input  : Image features $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$
         Metadata map $\mathbf{M} \in \mathbb{R}^{(N+2)\times H\times W}$
         Kernel size $K$ of DCNv2
         Modulation scale $k$
Output : Output features $\mathbf{Y} \in \mathbb{R}^{C'\times H\times W}$
Step 1: Offset branch
    $\Delta\mathbf{P}_{\text{img}} \leftarrow \text{Conv}(\mathbf{X}) \in \mathbb{R}^{2K^2\times H\times W}$
    $\Delta\mathbf{P}_{\text{meta}} \leftarrow \text{Conv}(\mathbf{M}) \in \mathbb{R}^{2K^2\times H\times W}$
    $\Delta\mathbf{P} \leftarrow \Delta\mathbf{P}_{\text{img}} + k\,\sigma(\Delta\mathbf{P}_{\text{meta}}) \in \mathbb{R}^{2K^2\times H\times W}$
Step 2: Mask branch
    $\mathbf{M}_{\text{mask}} \leftarrow 2\,\sigma(\text{Conv}(\mathbf{X})) \in \mathbb{R}^{K^2\times H\times W}$
Step 3: Deformable sampling
    $\mathbf{Y} \leftarrow \text{DCNv2}(\mathbf{X},\, \Delta\mathbf{P},\, \mathbf{M}_{\text{mask}})$        // see Eq. (1)
return $\mathbf{Y}$
```

3.3. Scale-Aware Spatial Feature Modulation

To provide spatially adaptive receptive fields for the detector, we adapt MDCN to bottleneck blocks of each CSPLayer in the feature pyramid network, which serves to enhance multi-scale representations. Specifically, we replace the second Convolution–BatchNorm–SiLU (CBS) block in each bottleneck (Figure 4(a)) with a deformable CBS block (DCBS) by substituting the convolution with DCNv2, resulting in the deformable bottleneck structure shown in Figure 4(b). This modification allows the network to capture spatially adaptive features via deformable operations with marginal computational cost.

We further extend this design to *Metadata-Guided Deformable CBS (MDCBS)* and *Metadata-Guided Deformable Bottleneck* (Figure 4(c)), where the sampling offsets are modulated by MDCN using platform metadata. This formulation enables the offsets to adapt jointly to image features and metadata priors, which we show in §4 to yield substantially larger performance gains compared to both the YOLOX baseline and its deformable bottleneck extension.

4. Experiments

We evaluate META-YOLO against other comparison models and show the effectiveness of our proposed approach. Specifically, our study investigates two main questions:

- **(RQ1)** Does incorporating auxiliary metadata consistently improve YOLOX and enable our models to outperform competitors, especially in lightweight settings? (§4.2)
- **(RQ2)** Is our feature extraction and modulation strategy more effective than existing alternatives? (§4.3)

4.1. Dataset

We present a custom EO aerial imagery dataset with frame-synchronized platform metadata. The dataset contains 62 flights and 327,023 annotated frames recorded at altitudes from 5 m to 550 m. Each frame is annotated with five vehicle categories: `car`, `bus`, `truck`, `utility vehicle (UV)`, and `transport vehicle (TV)`. The annotations include platform and sensor metadata fields following MISB ST 0601 [1], addressing the lack of public datasets with comprehensive, standard-compliant telemetry aligned with imagery. Appendix A presents further details.

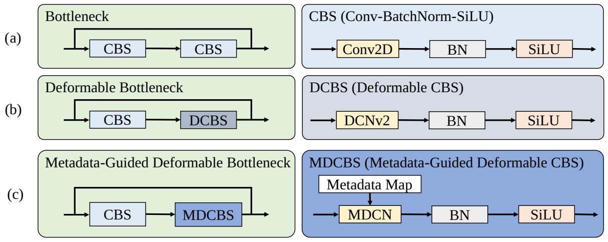

Figure 4. Comparison of bottleneck variants: (a) Baseline bottleneck and CBS blocks used in YOLOX [9], (b) Deformable bottleneck that applies spatially adaptive receptive fields using a DCBS block, (c) Our proposed Metadata-Guided Deformable Bottleneck, which further integrates platform metadata into MDCBS to generate adaptive offsets via MDCN.

4.2. Comparison with Lightweight Detectors

META-YOLO consistently outperforms existing lightweight detectors in both accuracy and efficiency. In particular, META-YOLO achieves higher AP under similar or lower GFLOPs, demonstrating its suitability for resource-limited aerial platforms. Table 1 summarizes the quantitative comparison of META-YOLO-Tiny and META-YOLO-S against their corresponding YOLO-series detectors.

<u>Comparison with YOLOX Baselines</u> META-YOLO models consistently improve upon the YOLOX baselines across lightweight scales [9]. They introduce only a marginal increase in computational cost.

- META-YOLO-Tiny achieves a 64.0 $\text{AP}^{Test}_{50:95}$, a substantial +8.7 point gain over YOLOX-Tiny's 55.3. This result comes with only +0.1M additional parameters and +0.8 GFLOPs. Notably, META-YOLO-Tiny surpasses the larger YOLOX-S model (62.4 $\text{AP}^{Test}_{50:95}$) while using only 57% of its parameters and 59% of its FLOPs.
- META-YOLO-S reaches 65.1 $\text{AP}^{Test}_{50:95}$, improving upon YOLOX-S (62.4) by +2.7 points with just +0.2M parameters and +1.0 GFLOPs.

These results strongly suggest that the additional metadata provides complementary cues that significantly boost the performance of lightweight backbones.

<u>Comparison with Alternative Detectors</u> We compare our models against recent lightweight detectors: YOLOv6 [14], YOLOv7 [32], YOLOv8 [10], and PP-YOLOE [40]. We use MMYOLO [4] to ensure consistent implementation.

Model	**#Epochs**	**#Params (M)**	**GFLOPs**	**AP**$^{Val}_{50:95}$	**AP**$^{Val}_{50}$	**AP**$^{Test}_{50:95}$	**AP**$^{Test}_{50}$	**AP**$^{Test}_{75}$	**AP**$^{Test}_{S}$	**AP**$^{Test}_{M}$	**AP**$^{Test}_{L}$
Nano & Tiny-sized Models											
YOLOX-Tiny [9]	300	5.0	18.2	53.4	**80.5**	55.3	76.6	64.0	36.3	62.3	76.5
YOLOv6-Nano [14]	400	4.3	13.2	57.5	79.2	61.5	83.1	71.6	42.1	68.7	87.0
YOLOv6-Tiny [14]	400	9.7	29.6	**59.2**	79.1	61.4	82.7	71.6	42.4	68.7	**89.5**
YOLOv7-Tiny [32]	300	6.0	15.8	33.9	62.1	35.9	63.4	38.5	17.7	43.3	57.5
YOLOv8-Nano [10]	500	3.0	9.8	55.1	74.3	59.6	78.9	69.8	39.7	66.7	74.7
META-YOLO-Tiny (Ours)	100	5.1	19.0	58.9	79.6	**64.0**	**85.1**	**74.2**	**42.5**	**71.2**	83.6
Small-sized Models											
YOLOX-S [9]	300	8.9	32.0	60.7	82.1	62.4	82.7	73.0	43.3	68.6	86.9
YOLOv6-S [14]	400	17.2	52.5	61.2	80.7	64.4	84.8	75.1	**45.9**	**71.7**	**88.8**
YOLOv8-S [10]	500	11.1	34.3	62.0	81.3	63.3	81.5	74.2	42.7	69.9	83.8
PP-YOLOE-S [40]	80	7.5	19.1	53.7	75.0	56.3	77.6	65.5	32.5	65.3	81.9
META-YOLO-S (Ours)	100	9.1	33.0	**62.3**	**83.2**	**65.1**	**85.6**	**75.8**	45.3	71.5	87.4

Table 1. Comparison of our proposed META-YOLO with lightweight detectors (Nano, Tiny, and Small). We report $AP_{50:95}$ and AP_{50} on both validation and test sets, and AP_{75}, AP_S, AP_M, and AP_L on the test sets. The bold values indicate the best performance in each column, and underlined values indicate the second-best. META-YOLO consistently outperforms YOLOX counterparts across lightweight scales, with only a marginal increase in parameters and FLOPs, while also showing competitive or superior performance against other detectors.

- META-YOLO-Tiny achieves the highest $AP^{Test}_{50:95}$ (64.0) among all nano and tiny-sized models, outperforming YOLOv6-Nano by +2.5 points and YOLOv6-Tiny by +2.6 points. It also outperforms YOLOv8-Nano by +4.4 points and demonstrates a massive improvement of +28.1 points over YOLOv7-Tiny.
- META-YOLO-S also secures the best $AP^{Test}_{50:95}$ (65.1) among all models. Moreover, it surpasses YOLOv6-S by +0.7 points, YOLOv8-S by +1.8 points, and PP-YOLOE-S by +8.8 points.

In summary, META-YOLO models significantly outperform their YOLOX baselines and establish a new state-of-the-art for lightweight object detection. They demonstrate superior performance across scales with competitive computational overhead.

4.3. Effectiveness of Metadata Integration

This section covers an ablation study, a hyperparameter analysis, a comparison with alternative modulation modules and feature extractors. All experiments in this section use the META-YOLO-Tiny model with a scaling factor $k = 2$ as the default setting.

Ablation Study We conduct a series of ablation studies to evaluate the contribution of each component in the META-YOLO architecture. Table 2 summarizes the results and reports both accuracy and computational cost. Starting from the YOLOX-Tiny baseline, which achieves 55.3 $AP^{Test}_{50:95}$ with 18.2 GFLOPs, we first add the DCNv2 as described in §3.3. This change alone increases accuracy by +6.0 points (+10.9%) to 61.3, while adding only +0.7 GFLOPs (+3.8%) in overhead. These results show that spatially adaptive receptive fields substantially improve feature representation with minimal computational expense.

Building on this result, we then incorporate the metadata-guided modulation module described in §3.2. This further increases the performance to 64.0 $AP^{Test}_{50:95}$. The metadata-guided module yields an additional gain of +2.7 points over the spatial-only variant, resulting in a total improvement of +8.7 points (+15.7%) over the original baseline. This substantial increase is achieved with only +0.1 GFLOPs additional cost. These results confirm that metadata guidance provides complementary contextual information beyond what visual data alone can offer. In short, the ablation study reveals two key findings:

- Scale-aware modulation provides significant geometry-aware improvements with minimal overhead.
- Metadata-guided modulation further enhances accuracy with negligible computational cost, improving performance while maintaining the detector's lightweight nature.

Ablation Settings	**AP**$^{Test}_{50:95}$	**GFLOPs**
YOLOX-Tiny	55.3	18.2
+ DCNv2 (Figure 4 (b))	61.3 (+10.9%)	18.9 (+3.8%)
+ MDCN (Figure 4 (c))	**64.0 (+15.7%)**	**19.0 (+4.4%)**

Table 2. Ablation study on metadata integration strategies and component-level variants.

Effect of Modulation Strength We investigate the effect of the scaling factor k, which controls the maximum modulation strength in Eq. 3. Figure 5 shows that all tested k values consistently improve performance over the YOLOX baseline (red line) and its deformable bottleneck extension (green line). These results demonstrate the robustness of the proposed modulation mechanism. However, performance does not monotonically increase as k becomes larger. For instance, small modulation ($k = 1$) yields limited gains, while excessively strong modulation ($k \geq 4$) causes performance degradation, likely due to unstable or over-amplified offset adjustments. The optimal trade-off is achieved at $k = 2$, where the modulation is strong enough to leverage metadata guidance while remaining stable across spatial locations. These findings highlight the importance of balancing modulation strength for effective feature adaptation.

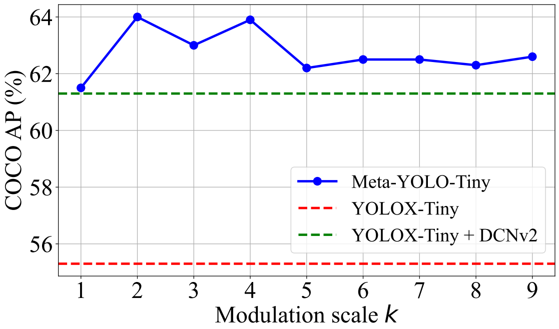

Figure 5. Performance analysis of the scale factor k, which controls the maximum modulation strength in Equation 3. All tested k values outperform the YOLOX [9] (red line) and its deformable bottleneck extension (green line). The best performance is achieved at $k = 2$.

Impact of Metadata on Modulation Methods To analyze the impact of metadata on different modulation methods, we applied it to Feature-wise Linear Modulation (FiLM) [22], a global modulation strategy. FiLM applies a global linear transformation to each feature channel, where modulation is realized by feature-wise scaling and shifting as:

$$\hat{f}_c = \gamma_c \cdot f_c + \beta_c$$

where f_c denotes the c-th feature channel, and γ_c, β_c are modulation parameters generated from conditioning inputs (*i.e*. metadata). This operation adjusts the feature map uniformly across all spatial locations, preventing it from adapting the modulation strength in a position-dependent manner.

As shown in Table 3, applying metadata with the global FiLM method, significantly boosts accuracy to 63.4 $\text{AP}^{Test}_{50:95}$ (+8.1, +14.7%), with a minimal increase of +0.6 GFLOPs (+3.3%). This result demonstrates that metadata is highly effective even with a simple, global modulation strategy. Furthermore, our proposed deformable convolution-based modulation maximizes the potential of metadata through spatially-adaptive adjustments. This approach further elevates performance to 64.0 $\text{AP}^{Test}_{50:95}$ (+8.7, +15.7% over the baseline), with a comparable overhead of +0.8 GFLOPs (+4.4%). In conclusion, our findings confirm that while platform metadata itself serves as a powerful factor for performance improvement of lightweight detectors, its impact is maximized when paired with a spatially-aware mechanism that directly adapts sampling positions.

Effect of Feature Extractor As discussed in §3.2, we construct a metadata map and process it through a convolutional module to generate feature for direct fusion with visual features. We compare this approach with two alternative extractors: a residual MLP [20] and a dynamic MLP [42].

Modulation Method	**$\text{AP}^{Test}_{50:95}$**	**GFLOPs**
None (YOLOX-Tiny)	55.3	18.2
FiLM [22]	63.4 (+14.6%)	18.8 (+3.3%)
MDCN (Ours)	**64.0 (+15.7%)**	**19.0 (+4.4%)**

Table 3. Comparison of modulation methods. Applying metadata through a global strategy (FiLM) improves accuracy with minimal overhead, confirming the effectiveness of metadata. Our spatially-adaptive MDCN further maximizes this benefit.

Feature Extractor	**$\text{AP}^{Test}_{50:95}$**	**GFLOPs**
None (YOLOX-Tiny)	55.3	18.2
Residual MLP [20]	61.4 (+11.0%)	19.0 (+4.4%)
Dynamic MLP [42]	63.2 (+14.3%)	19.0 (+4.4%)
Metadata Map (Ours)	**64.0 (+15.7%)**	**19.0 (+4.4%)**

Table 4. Comparison of metadata feature extractors. The Metadata Map extractor achieves the highest accuracy with similar overhead.

Table 4 reports the results. The residual MLP improves performance to 61.4 $\text{AP}^{Test}_{50:95}$ (+6.1, +11.0%) but increases complexity to 19.0 GFLOPs. The dynamic MLP further enhances accuracy to 63.2 $\text{AP}^{Test}_{50:95}$ (+7.9, +14.3%), highlighting the benefit of adapting metadata transformation to input variation. The metadata map extractor achieves the best result, 64.0 $\text{AP}^{Test}_{50:95}$ (+8.7, +15.7% over baseline), while keeping computational cost identical to the MLP-based alternatives (19.0 GFLOPs). These results highlight that convolutional processing of metadata maps is more effective than vector-based MLP encoders. Unlike MLPs, which treat metadata as flat vectors, the metadata map preserves spatial alignment and allows convolutional filters to capture local correlations between metadata channels. This design provides stronger accuracy gains at no extra cost, demonstrating the effectiveness of metadata map features for guiding deformable modulation.

4.4. Visualization of Predictions

We present qualitative detection results of META-YOLO-Tiny on our dataset. As shown in Figure 6, the detector successfully localizes vehicles across diverse environments, including highways, dirt roads, residential areas, and parking lots. The results demonstrate that META-YOLO-Tiny maintains stable performance under varying scene contexts and imaging conditions, such as oblique viewpoints, different altitudes, and cluttered backgrounds. These examples highlight the practical robustness of metadata-guided modulation in real-world aerial scenarios.

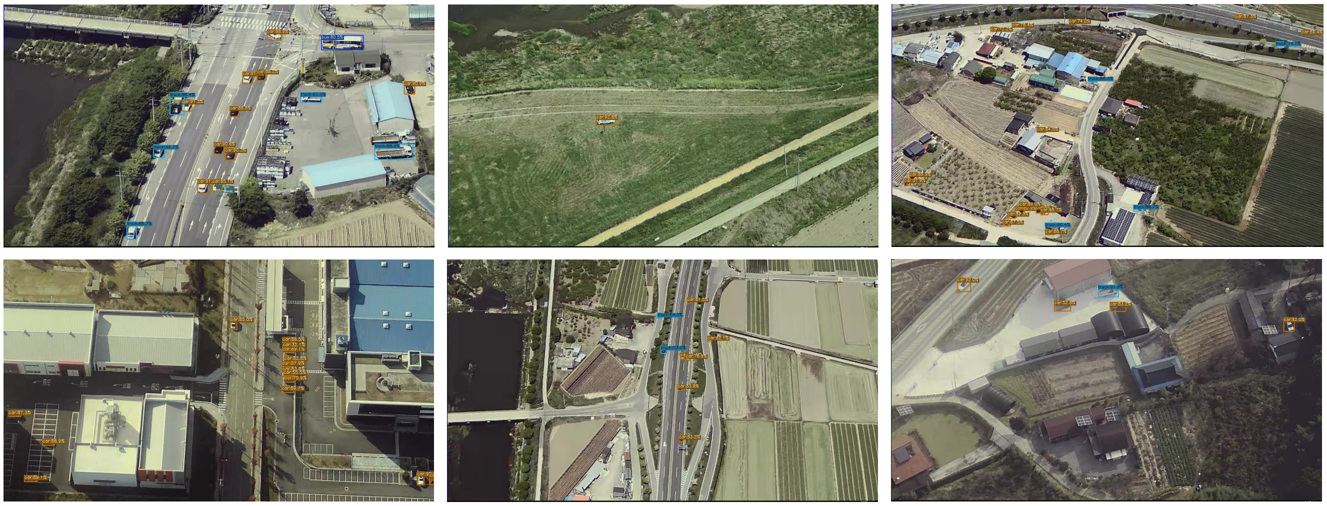

Figure 6. Qualitative detection results of META-YOLO-Tiny across diverse environments, including highways, dirt roads, residential areas, and parking lots. These examples illustrate the robustness of our detector under varying scene contexts and imaging conditions.

5. Discussion

Generalizability Beyond Custom Dataset Our method generalizes across platforms by leveraging geometric cues from telemetry, not dataset-specific features. We parse and min–max normalize MISB ST 0601 [1] fields, injecting them into deformable layers to ensure compatibility with sensors and airframes using this standard. The metadata map aligns with image coordinates and adapts to feature map resolution, enabling deployment across cameras and input sizes. Because modulation acts on sampling locations rather than class logits, the mechanism remains category agnostic and less sensitive to label shifts. To address variation in appearance and metadata, we bound metadata influence using per-channel normalization and a scaled activation to limit offset magnitude, maintaining stability without visual processing.

Extending Other Models with Metadata Our approach is broadly compatible with modern object detectors, including architectures beyond those evaluated in this work. Since we focus on models implemented in MMYOLO [4], recent YOLO-series detectors [13, 31], and real-time DETR-series detectors [19, 34, 46] are not included in our evaluation. Nevertheless, our mechanism learns sampling locations without changing the core detector architecture. This makes our method applicable to the latest detectors and particularly suitable for end-to-end models where deformable sampling plays a central role, such as Deformable DETR [48] and DINO [45]. Consequently, we expect the largest performance gains in frameworks that govern receptive fields entirely through deformable offsets.

Future Work First, we plan to study cross-dataset transfer without fine-tuning. We also evaluate a no metadata mode by freezing the offset branch, which isolates the visual baseline. This quantifies generalization under appearance shift and missing or delayed fields. Per field ablations will identify critical signals and set fallback thresholds, informing calibration for new platforms.

Second, we propose extending metadata-guided modulation to the recent detectors to show our approach is model-agnostic. We do not evaluate this in the current paper, but this extension tests model-agnostic claims and targets strong scale variation. Expected outcomes are improved small object accuracy and less reliance on multi scale training. Minor schedule tuning may be needed, with minimal latency impact since the branch reuses backbone features.

6. Related work

Recent advances in deep learning have significantly improved general-purpose object detection [9, 10, 25, 29, 32]. Despite these gains, object detection in aerial imagery, especially from drones or UAVs, remains difficult due to small and variably sized objects, viewpoint variations, and strict real-time and resource limits.

Aerial object detection research divides into several main categories. First, some studies adapt general-purpose detectors to aerial imagery. These works focus on small object detection, lightweight model design, and domain adaptation [5, 35, 36, 39]. Second, other research develops aerial-specific techniques such as oriented object detectors [11, 38]. Third, a category exploits auxiliary information, such as platform metadata, sensor data, and multi-modal contextual cues, to guide detection [2, 16, 36, 43]. Benchmark datasets and standardized protocols also shape the aerial imagery research landscape, including DOTA, VisDrone, UAVDT, and AI-TOD [6, 7, 33, 37]. This work follows the third category. It leverages the metadata to improve detection performance.

Approximate Metadata from Visual Features Early studies in this area integrate cues such as flight parameters into the detection process. To exploit these cues, one strategy trains detectors jointly with auxiliary prediction tasks to enhance feature representations [16, 36, 43]. For example, GSDet introduces a GSD identification subnet to refine scale reasoning for regions of interest. It uses the resulting GSD distribution with each RoI's size to infer physical extent and re-weight RoI features [16]. GSDDet [43] extracts GSD-aware features using a classification network and combines them with an attention framework for object detection. These methods approximate metadata from visual features rather than directly exploiting metadata when it is available. In contrast, NDFT [36] uses flight metadata, such as altitude, viewing angle, and weather condition, as nuisance labels. It applies adversarial training to separate these labels from shared features and improve condition robustness. However, NDFT treats metadata only as a source of global invariance and does not leverage telemetry as a positive cue.

Directly Integrate Physical Metadata Another approach directly integrates non-visual platform metadata into detection networks using conditioning or normalization [17, 21, 27]. Adaptive Resizing uses metadata-informed scale normalization to address scale variance [21]. AuxDet introduces an MLP-based multi-modal modulation module that combines auxiliary metadata, such as sensor platform, band type, and image resolution, with visual features [27]. AltiDet combines UAV altitude with a fusion pyramid to improve tiny-object detection [17]. Most of these studies rely on a limited subset of metadata, focusing mainly on altitude or categorical descriptors. In contrast, our approach exploits a comprehensive set of platform and sensor fields defined in MISB ST 0601 [1]. Beyond detection models, AU-AIR represents a public multi-modal UAV dataset that pairs RGB imagery with synchronized flight telemetry, including inertial measurement units (IMU), GPS, and altitude [2]. However, AU-AIR was collected at altitudes below 30 m for traffic surveillance, thereby restricting its utility to near-ground UAV operations and limiting generalization to higher-altitude aerial imagery. As a result, targets occupy relatively large pixel areas and the dataset under-represents small-sized object scenarios.

Platform Metadata in Visual Recognition Platform and sensor metadata also benefit other vision tasks. In fine-grained image classification (*e.g.* bird species), geolocation and observation time from platform metadata help improve accuracy [20, 42]. Visually similar species can be difficult to distinguish, but their geographical ranges or seasonal migratory patterns provide critical cues. In video stabilization, metadata from IMU, such as gyroscope and accelerometer readings, is leveraged to estimate the camera's motion [15, 28, 44]. By integrating these sensor signals with visual cues, a system can more accurately recover the camera pose and apply corrective geometric transformations to each frame, thereby compensating for platform-induced motion and producing stabilized video sequences. Building on this, visual–inertial Simultaneous Localization and Mapping (VI-SLAM), which is a technique that jointly estimates camera trajectory and reconstructs a map of the environment, fuses camera imagery with inertial sensor readings (gyroscope, accelerometer) to provide robust and drift-reduced state estimation [3, 26]. Lastly, in multi-object tracking, platform and sensor metadata are commonly exploited to decouple camera ego-motion from object motion, enabling ego-motion compensation that stabilizes target trajectories [12, 41]. Such metadata-guided world-coordinate reasoning improves long-term ID preservation and robustness under occlusions or rapid maneuvers.

Summary of Contribution While previous work has often approximated metadata from visual features or relied on a limited subset of non-spatial cues, we demonstrate the effectiveness of comprehensive standardized telemetry and introduce a novel modulation module that leverages metadata to adaptively modulate spatial features. This provides a physically grounded mechanism for adapting the model's receptive field to object scale variations in real time, offering a robust paradigm for metadata-guided aerial object detection.

7. Conclusion

In this paper, we propose Meta-YOLO, a real-time aerial object detector that leverages platform metadata to guide scale-aware spatial feature extraction. Meta-YOLO injects normalized telemetry into deformable convolution layers, allowing receptive fields to adapt dynamically to object scale and scene geometry. Our metadata-guided modulation and scale-aware feature adaptation address key limitations of conventional detectors, which often ignore contextual cues and struggle with scale variation in cluttered aerial imagery. Built on YOLOX, Meta-YOLO demonstrates consistent gains over strong baselines across a large-scale benchmark of 327K annotated aerial frames. Extensive experiments validate that Meta-YOLO improves detection accuracy by up to +8.7 points in AP, while preserving real-time performance and low computational cost.

Acknowledgments. This work was supported by the Agency for Defense Development Grant funded by the Korean government. We would also like to thank Yong-Duk Kim, our project leader, for his continuous support and guidance throughout this project.

References

[1] Misb st 0601.8: Uas datalink local set. Technical report, Motion Imagery Standards Board (MISB), 2014. 1, 2, 4, 7, 8

[2] Ilker Bozcan and Erdal Kayacan. Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8504–8510. IEEE, 2020. 7, 8

[3] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE transactions on robotics*, 37(6):1874–1890, 2021. 8

[4] MMYOLO Contributors. MMYOLO: OpenMMLab YOLO series toolbox and benchmark. https://github.com/open-mmlab/mmyolo, 2022. 4, 7, 11

[5] Lixia Deng, Lingyun Bi, Hongquan Li, Haonan Chen, Xuehu Duan, Haitong Lou, Hongyu Zhang, Jingxue Bi, and Haiying Liu. Lightweight aerial image object detection algorithm based on improved yolov5s. *Scientific reports*, 13(1):7817, 2023. 7

[6] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–386, 2018. 7

[7] Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Lin, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, et al. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 7

[8] Beatriz Felipe-García, David Hernández-López, and José Luis Lerma. Analysis of the ground sample distance on large photogrammetric surveys. *Applied Geomatics*, 4(4):231–244, 2012. 1

[9] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 1, 3, 4, 5, 6, 7, 11, 12

[10] Jocher Glenn. Yolov8. *https://github.com/ultralytics/ultralytics/tree/main*, 2023. 4, 5, 7, 12

[11] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2786–2795, 2021. 7

[12] Hung-Min Hsu, Jiarui Cai, Yizhou Wang, Jenq-Neng Hwang, and Kwang-Ju Kim. Multi-target multi-camera tracking of vehicles using metadata-aided re-id and trajectory-based camera link model. *IEEE Transactions on Image Processing*, 30: 5198–5210, 2021. 8

[13] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. 7

[14] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022. 4, 5

[15] Chen Li, Li Song, Shuai Chen, Rong Xie, and Wenjun Zhang. Deep online video stabilization using imu sensors. *IEEE Transactions on Multimedia*, 25:2047–2060, 2022. 8

[16] Wei Li, Wei Wei, and Lei Zhang. Gsdet: Object detection in aerial images based on scale reasoning. *IEEE Transactions on Image Processing*, 30:4599–4609, 2021. 1, 7, 8

[17] Chan Yue Liew, Joanne Mun-Yee Lim, Chee Pin Tan, and Raja Mazhar Mohar Bin Tun Mohar. Altitude-informed fusion pyramid network for multi-scale waste detection in unmanned aerial vehicle images. *Engineering Applications of Artificial Intelligence*, 153:110814, 2025. 1, 8

[18] 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*, pages 740–755. Springer, 2014. 11

[19] Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer. *arXiv preprint arXiv:2407.17140*, 2024. 7

[20] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9596–9606, 2019. 6, 8

[21] Martin Messmer, Benjamin Kiefer, and Andreas Zell. Gaining scale invariance in uav bird's eye view object detection by adaptive resizing. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3588–3594. IEEE, 2022. 2, 8

[22] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 6

[23] Yalong Pi, Nipun D Nath, and Amir H Behzadan. Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. *Advanced Engineering Informatics*, 43:101009, 2020. 1

[24] Aref Miri Rekavandi, Lian Xu, Farid Boussaid, Abd-Krim Seghouane, Stephen Hoefs, and Mohammed Bennamoun. A guide to image-and video-based small object detection using deep learning: case study of maritime surveillance. *IEEE Transactions on Intelligent Transportation Systems*, 2025. 1

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 1, 7, 12

[26] Myriam Servières, Valérie Renaudin, Alexis Dupuis, and Nicolas Antigny. Visual and visual-inertial slam: State of the art, classification, and experimental benchmarking. *Journal of Sensors*, 2021(1):2054828, 2021. 8

[27] Yangting Shi, Renjie He, Le Hui, Xiang Li, Jian Yang, Ming-Ming Cheng, and Yimian Dai. Auxdet: Auxiliary metadata matters for omni-domain infrared small target detection. *arXiv preprint arXiv:2505.15184*, 2025. 2, 8

[28] Zhenmei Shi, Fuhao Shi, Wei-Sheng Lai, Chia-Kai Liang, and Yingyu Liang. Deep online fused video stabilization. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1250–1258, 2022. 8

[29] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021. 7, 12

[30] Wei Sun, Liang Dai, Xiaorui Zhang, Pengshuai Chang, and Xiaozheng He. Rsod: Real-time small object detection algorithm in uav-based traffic monitoring. *Applied Intelligence*, 52(8):8448–8463, 2022. 1

[31] Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025. 7

[32] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023. 4, 5, 7, 12

[33] Jinwang Wang, Wen Yang, Haowen Guo, Ruixiang Zhang, and Gui-Song Xia. Tiny object detection in aerial images. In *2020 25th international conference on pattern recognition (ICPR)*, pages 3791–3798. IEEE, 2021. 7

[34] Shuo Wang, Chunlong Xia, Feng Lv, and Yifeng Shi. Rt-detrv3: Real-time end-to-end object detection with hierarchical dense positive supervision. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1628–1636. IEEE, 2025. 7

[35] Xin Wang, Ning He, Chen Hong, Fengxi Sun, Wenjing Han, and Qi Wang. Yolo-erf: lightweight object detector for uav aerial images. *Multimedia Systems*, 29(6):3329–3339, 2023. 7

[36] Zhenyu Wu, Karthik Suresh, Priya Narayanan, Hongyu Xu, Heesung Kwon, and Zhangyang Wang. Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1201–1210, 2019. 7, 8

[37] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 7

[38] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3520–3529, 2021. 7

[39] Chang Xu, Jinwang Wang, Wen Yang, and Lei Yu. Dot distance for tiny object detection in aerial images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1192–1201, 2021. 7

[40] Shangliang Xu, Xinxin Wang, Wenyu Lv, Qinyao Chang, Cheng Cui, Kaipeng Deng, Guanzhong Wang, Qingqing Dang, Shengyun Wei, Yuning Du, and Baohua Lai. Pp-yoloe: An evolved version of yolo. *ArXiv*, abs/2203.16250, 2022. 4, 5, 12

[41] Cheng-Yen Yang, Hsiang-Wei Huang, Zhongyu Jiang, Heng-Cheng Kuo, Jie Mei, Chung-I Huang, and Jenq-Neng Hwang. Sea you later: Metadata-guided long-term re-identification for uav-based multi-object tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 805–812, 2024. 8

[42] Lingfeng Yang, Xiang Li, Renjie Song, Borui Zhao, Juntian Tao, Shihao Zhou, Jiajun Liang, and Jian Yang. Dynamic mlp for fine-grained image classification by leveraging geographical and temporal information. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10945–10954, 2022. 6, 8

[43] Yunuo Yang, Cheng Wang, Zhipeng Cai, Pinqing Song, Guanjie Huang, Ming Cheng, and Yu Zang. Gsddet: Ground sample distance-guided object detection for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–12, 2023. 1, 7, 8

[44] Jiyang Yu, Tianhao Zhang, Fuhao Shi Google, Lei He, and Chia-Kai Liang. Sensorflow: Sensor and image fused video stabilization. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8454–8463. IEEE, 2025. 8

[45] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 7

[46] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16965–16974, 2024. 7

[47] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019. 2

[48] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1, 7

A. Experimental Setup

Dataset As we described in §4.1, we construct a custom dataset using drones equipped with EO cameras. The dataset contains videos with synchronized images and metadata collected from altitudes between 5 m and 550 m. Each video is recorded at 30 frames per second (fps) with 1280×720 resolution. We allocate 240,045 images (73%) for training, 46,818 images (14%) for validation, and 40,160 images (12%) for testing. Frames from the same video remain in the same subset to prevent data leakage. From the training set, we sample 1 out of every 10 frames to reduce redundancy among adjacent frames. Table 6 summarizes the dataset statistics.

Experimental Environment We evaluate performance with the standard COCO metrics [18]. These metrics include Average Precision (AP) and its variants. $\mathbf{AP}^{\text{val}}_{50:95}$ and $\mathbf{AP}^{\text{test}}_{50:95}$ measure average precision across IoU thresholds from 0.50 to 0.95 in steps of 0.05 on both the validation and test datasets, respectively. We also report $\mathbf{AP}^{\text{test}}_{50}$ and $\mathbf{AP}^{\text{test}}_{75}$ as IoU-specific scores, and $\mathbf{AP}^{\text{test}}_{S}$, $\mathbf{AP}^{\text{test}}_{M}$, and $\mathbf{AP}^{\text{test}}_{L}$ as scale-specific scores on the test dataset.

Implementation Details We start from YOLOX [9] pretrained on the COCO dataset [18], and fine-tune it on our dataset. We train META-YOLO with the SGD optimizer using eight NVIDIA A100 GPUs (40GB) with a batch size of 64 per GPU. We set the basic learning rate to 0.01 and the weight decay to 0.0005. We apply the cosine learning rate schedule of [9] and the exponential moving average (EMA) with ema_decay of 0.999. During training, we apply HSV augmentation and random flip operations. The main hyperparameters of META-YOLO are listed in Table 5 (refer to META-YOLO-Tiny for detailed configuration).

Comparison Model Settings For all baselines, the input resolution is fixed to 1280×768. To satisfy the 2^n resolution constraint of modern detectors, we apply minimal padding along the height dimension. All comparison models are trained using the MMYOLO framework [4] and initialized from the COCO-pretrained weights provided by the corresponding implementations. We adopt the default training strategy of MMYOLO, including the number of epochs, learning rate schedule, and data augmentations.

Item	**Value**
input size	(1280, 720)
activation function	silu
depth	0.33
width	0.375
scheduler	SGD
basic learning rate	0.01
weight decay	0.0005
cosine learning rate schedule	True
momentum	0.9
ema decay	0.999
flip probability	0.5
maximum epoch	100
test confidence	0.001
nms threshold	0.65
number of metadata	7
metadata strength	2

Table 5. Main Hyperparameters of META-YOLO-Tiny

B. Comparison with High-Capacity Models

To evaluate the scalability of our approach beyond the lightweight regime, we extend our experiments to large-capacity models. Specifically, we compare META-YOLO-L and META-YOLO-X with their YOLOX counterparts, other state-of-the-art YOLO variants, and prominent two-stage detectors such as Faster R-CNN and Sparse R-CNN. The results are summarized in Table 7.

META-YOLO maintains consistent improvements over the YOLOX baselines, even at a large scale. For instance, META-YOLO-L achieves 68.4 AP$^{Test}_{50:95}$, a notable +2.2 point gain over YOLOX-L. Similarly, META-YOLO-X reaches 68.9 AP$^{Test}_{50:95}$, outperforming YOLOX-X by +1.8 points. These improvements also extend to finer metrics; for example, Meta-YOLO-L shows enhanced performance in AP$_{75}$ (79.5 vs. 77.4) and AP$_{S}$ (50.8 vs. 47.1), indicating that our approach continues to benefit localization precision and small-object detection in high-capacity settings.

In comparison with other advanced detectors, META-YOLO demonstrates strong generalization and competitive performance on the test set. While models like YOLOv8-L show higher accuracy on the validation set, META-YOLO-L ultimately achieves superior performance on the test set with 68.4AP$^{Test}_{50:95}$ compared to YOLOv8-L's 67.9, and does so with slightly fewer GFLOPs (191.6 vs. 198.0). Furthermore, our model significantly outperforms other efficient detectors like PP-YOLOE, with META-YOLO-L surpassing PP-YOLOE-L by +1.6 points. A similar trend is observed in the XLarge scale, where META-YOLO-X substantially exceeds PP-YOLOE-X by +3.4 points. While YOLOv8-X holds a slight edge in overall AP$^{Test}_{50:95}$, our META-YOLO-X excels in crucial metrics, achieving a state-of-the-art 88.1 AP$^{Test}_{50}$ and demonstrating better performance on small objects AP$^{Test}_{S}$ When compared against two-stage detectors, both META-YOLO-L and META-YOLO-X provide substantially higher accuracy than models like Sparse R-CNN while maintaining the efficiency inherent in one-stage designs.

Overall, these results confirm that the proposed metadata-guided modulation scales robustly to high-capacity models. It is worth noting, however, that the relative performance gains are more pronounced in the lightweight regime. This suggests that the benefits of metadata are most significant when a model's intrinsic representational capacity is constrained, offering a compelling direction for future research on efficient model design.

Split	**# Video Sequence**	**# Image**	**# Car**	**# Bus**	**# Truck**	**# UV**	**# TV**
Train	44	240,045 (73%)	992,518 (75%)	12,041 (77%)	371,162 (79%)	91,290 (77%)	91,348 (74%)
Valid	9	46,818 (14%)	145,488 (11%)	1,929 (12%)	49,514 (10%)	14,597 (12%)	16,145 (13%)
Test	9	40,160 (12%)	192,896 (14%)	1,672 (11%)	51,359 (11%)	12,784 (11%)	15,734 (13%)
Total	62	327,023	1,330,902	15,642	472,035	118,671	123,227

Table 6. Statistics of the dataset. The dataset consists of 62 video sequences, with 327,023 frames across 5 categories: car, bus, truck, utility vehicle, and transport vehicle. We split the dataset at the video-level and sample 10% of frames for training after the split.

Model	**#Epochs**	**#Params (M)**	**GFLOPs**	$\textbf{AP}^{Val}_{50:95}$	$\textbf{AP}^{Val}_{50}$	$\textbf{AP}^{Test}_{50:95}$	$\textbf{AP}^{Test}_{50}$	$\textbf{AP}^{Test}_{75}$	$\textbf{AP}^{Test}_{S}$	$\textbf{AP}^{Test}_{M}$	$\textbf{AP}^{Test}_{L}$
Two-stage Detectors											
Faster R-CNN (R50) [25]	24	41.4	191.0	57.2	79.9	57.1	79.5	67.3	36.3	65.0	85.5
Faster R-CNN (R101) [25]	24	60.4	261.0	55.7	78.4	52.3	73.4	62.0	32.1	60.1	85.8
Sparse R-CNN (R50) [29]	36	106.0	158.0	**58.1**	80.6	**60.2**	**84.7**	**70.7**	**39.8**	66.9	87.0
Sparse R-CNN (R101) [29]	36	125.0	227.0	57.9	**80.7**	60.0	84.5	70.4	37.3	**67.2**	**87.9**
Large-sized Models											
YOLOX-L [9]	300	54.2	186.0	64.9	84.6	66.2	84.7	77.4	47.1	72.3	85.1
YOLOv7-L [32]	300	36.5	124.0	38.0	66.0	44.1	71.7	49.8	25.9	50.0	56.7
YOLOv8-L [10]	500	43.6	198.0	**69.0**	**87.1**	67.9	85.3	79.4	46.5	**74.4**	80.4
PP-YOLOE-L [40]	80	51.3	129.0	65.9	85.5	66.8	87.4	78.5	43.0	73.7	89.2
Meta-YOLO-L	100	55.0	191.6	66.3	85.7	**68.4**	**87.6**	**79.5**	**50.8**	74.0	**90.3**
XLarge-sized Models											
YOLOX-X [9]	300	99.0	338.0	66.0	85.7	67.1	85.1	78.9	47.6	73.4	**90.3**
YOLOv7-X [32]	300	70.8	226.0	42.0	64.9	47.9	71.4	55.9	27.0	54.9	57.8
YOLOv8-X [10]	500	68.2	309.0	**69.2**	**86.8**	**69.2**	86.7	**80.4**	50.5	**74.9**	84.3
PP-YOLOE-X [40]	80	97.3	244.0	64.2	82.2	65.5	84.0	76.9	45.4	71.8	87.4
Meta-YOLO-X	100	100.4	346.3	67.7	86.3	68.9	**88.1**	80.2	**51.0**	74.4	88.6

Table 7. Comparison of our proposed META-YOLO with large-capacity detectors. We report results of META-YOLO-L and META-YOLO-X against YOLOX, YOLOv7, YOLOv8, PP-YOLOE, and two-stage counterparts.