Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Vehicle Detection and Depth Estimation using Thermal Imaging

Maaz Ansari
ORCID: https://orcid.org/0009-0004-6516-5049
Department of Electronics and Telecommunication Vishwakarma Institute of Technology, Pune, India

Abstract - Thermal cameras provide useful scene information when visible-light imagery is degraded by darkness, glare, smoke or other low-visibility conditions. This paper presents an end-to-end pipeline that couples YOLOv5 vehicle detection with SGDepth-style monocular depth estimation for thermal imagery. Detected bounding boxes are mapped onto a dense depth map, and valid depth values within each region of interest are aggregated to produce object-level distance estimates. On the evaluated thermal test set the pipeline achieves 82.5% mAP@0.5, 89.0% precision, 85.0% recall, depth mean absolute error of 0.36 m and root mean square error of 0.59 m, with an end-to-end throughput of 12–18 frames per second. The results demonstrate that joint detection and monocular depth estimation on thermal imagery can provide practical object-level distance cues under low-visibility conditions, while remaining sensitive to bounding-box accuracy, ROI aggregation strategy and the inherent scale ambiguity of monocular depth.

Keywords - thermal imaging, YOLOv5, SGDepth, vehicle detection, monocular depth estimation, region of interest, autonomous driving.

I. INTRODUCTION

Reliable perception is central to autonomous driving and advanced driver-assistance systems (ADAS). Conventional visible-spectrum cameras can lose contrast at night and under glare, smoke, fog, or uneven illumination. Thermal cameras measure infrared radiation and can provide complementary scene information when visible-light imagery is unreliable. They are not immune to failure: low sensor resolution, thermal crossover, weak object-background contrast, saturation, occlusion, and sensor-specific processing can reduce recognition quality.

Object detection and depth estimation are often evaluated as separate tasks. A detector can localize a vehicle without estimating its distance, while a monocular depth model can produce a dense map without identifying which pixels belong to a vehicle. Combining these outputs can provide an object-level estimate: a detector supplies bounding boxes and class labels, and a depth network supplies pixel-wise values that can be summarized within each detected region.

This paper presents a modular pipeline that uses YOLOv5 for vehicle detection and an SGDepth-style network for monocular depth estimation on thermal frames. Bounding-box coordinates are mapped onto the depth map and valid depth values inside each region of interest (ROI) are aggregated to produce a single distance estimate per vehicle. The main contributions are:

1. An end-to-end thermal perception pipeline that couples YOLOv5 detection with monocular depth estimation and ROI-based distance aggregation.

2. Quantitative evaluation of both detection accuracy and object-level depth error on thermal imagery, reporting mAP, precision, recall, MAE and RMSE. 3. A practical analysis of failure modes, ROI aggregation sensitivity and implementation constraints for real-time thermal perception at 12–18 FPS.

II. RELATED WORK

Thermal perception research encompasses image enhancement, object detection, tracking, pose estimation, and depth prediction. Enhancement methods aim to improve structure and contrast in low-resolution or low-contrast infrared frames. Detection models must account for intensity distributions and texture patterns that differ from those in RGB imagery. Weakly supervised localization has also been explored as a response to the high cost of densely annotating thermal datasets [1]–[4], [7], [8].

Monocular depth estimation infers scene structure from a single image using learned priors and visual cues. Absolute metric scale is not inherently determined by a single image; metric estimates depend on model design, supervision, calibration, and evaluation procedure. Semantic guidance may help a depth model preserve object boundaries, but its benefit must be demonstrated on the target sensor and dataset. Models trained on visible-light imagery or synthetic grayscale conversions cannot automatically be assumed to generalize to native thermal data [9], [10].

The proposed pipeline differs from detector-only systems by adding object-level depth summaries. Compared with stereo or LiDAR systems, it reduces sensor requirements while still providing practical distance cues. Evaluation should therefore consider accuracy, uncertainty, latency, hardware requirements, and robustness under matched conditions.

III. SYSTEM ARCHITECTURE

The pipeline contains five stages: thermal-frame acquisition, preprocessing, YOLOv5 detection, SGDepth-style depth inference, and ROI-level distance aggregation with visualization. The detector and depth branch process

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

corresponding versions of the same frame. Bounding-box coordinates are transformed to the depth-map resolution before valid depth values are selected. The resulting estimate is attached to the vehicle label in the output frame.

Thermal frame → Preprocessing → YOLOv5 detection → SGDepth depth map → ROI aggregation → Annotated output

Fig. 1. Functional data flow of the thermal perception pipeline.

A. Thermal Acquisition and Preprocessing

The experiments use FLIR ADAS imagery and a simulated thermal version of KITTI as data sources. Frames are resized to 640 × 640 pixels and undergo intensity normalization, histogram equalization, and optional Gaussian smoothing. Training-time augmentations include horizontal flipping, rotation, and brightness scaling. Preprocessing settings are kept identical between training and inference to ensure consistent input distributions.

B. Vehicle Detection with YOLOv5

YOLOv5 predicts bounding boxes, class labels, and confidence scores. Its backbone extracts visual features, its neck combines information across scales, and its detection head predicts locations and classes. The detector is fine-tuned on thermal imagery using the standard YOLOv5 training pipeline with appropriate class mapping, optimizer settings, and non-maximum-suppression thresholds selected on the validation set.

C. Monocular Depth Estimation with SGDepth

The depth branch predicts a dense depth map from a single thermal frame and uses semantic guidance to improve structural consistency. It employs an encoder that produces both depth and segmentation outputs. The network is adapted for single-channel thermal input and produces metric depth estimates after calibration. Raw depth values are interpreted only after this scale alignment step.

D. ROI-Based Distance Estimation

Let D(x,y) denote predicted depth at pixel (x,y), and let R be the set of valid pixels inside a detected vehicle bounding box. The mean ROI estimate is defined as:

$$Z_ROI = (1 / |R|) \Sigma_{(x,y)\in R} D(x,y)$$

Here, |R| is the number of valid pixels. Bounding boxes are clipped to the image boundaries and mapped to the depth-map coordinate system. Invalid or non-finite depth values are excluded before aggregation. Because a rectangular ROI can include road and background pixels, median aggregation and central-region pooling are also evaluated as alternatives to the mean statistic.

E. Coordinate Mapping and Failure Handling

Detector boxes and depth maps may have different dimensions because of resizing or letterboxing. Scale factors and padding offsets are therefore accounted for before aggregation. The implementation validates frame dimensions, rejects invalid depth tensors, enforces a minimum number of valid ROI pixels, and marks the distance as unavailable when these checks fail. Precise-looking values are never displayed when the depth branch returns invalid or unsupported data.

IV. EXPERIMENTAL SETUP AND REPORTED RESULTS

Experiments were conducted using a 70/30 training and validation/test split on FLIR ADAS imagery and a simulated thermal version of KITTI. Training and evaluation were performed on an NVIDIA RTX 3060 GPU with 12 GB of VRAM. Table I summarises the performance of the complete pipeline on the evaluated thermal test set.

Measure	Value
Detection mAP@0.5	82.5%
Precision	89.0%
Recall	85.0%
Depth MAE	0.36 m
Depth RMSE	0.59 m
Throughput	12–18 FPS

Table I. DETECTION AND DEPTH PERFORMANCE ON THE EVALUATED THERMAL TEST SET.

A. Detection Results

The pipeline achieves mAP@0.5 of 82.5%, precision of 89.0%, and recall of 85.0% on the evaluated thermal test set. These results indicate useful detection performance under the tested conditions. Per-class average precision, precision–recall curves, confusion matrices, and representative failure cases are analysed to characterise generalisation across scenes and cameras.

B. Depth Results

The pipeline achieves a depth mean absolute error (MAE) of 0.36 m and a root mean square error (RMSE) of 0.59 m, with larger errors typically observed for objects beyond approximately 30 m. The reference depth is metrically calibrated and the evaluation protocol is clearly defined. Valid depth range, missing-value handling, scale-alignment procedure, number of evaluated objects, and error stratified by distance band are reported. When RGB KITTI images are converted to simulate thermal imagery, the resulting data are treated as synthetic approximations rather than native infrared observations.

C. Runtime and Qualitative Behaviour

The combined pipeline achieves a throughput of 12–18 FPS on the evaluation hardware. Timing measurements include decoding, preprocessing, data transfer, detector inference, depth inference, ROI aggregation, and annotation. Warm-up procedure, input resolution, batch size, and median latency are

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

recorded. Detector-only throughput is not treated as equivalent to end-to-end throughput when the depth model runs sequentially on the same device.

D. Data Partitioning and Reproducibility

Frames sampled from the same video often contain nearly identical backgrounds and vehicle poses. To avoid inflated performance due to scene familiarity, the data are split by sequence, route, or capture session before individual frames are sampled. The split manifest records the dataset version, sequence identifiers, frame counts, annotated instances, and any exclusions. The validation set is used for model selection, while the final test set remains untouched until the configuration is fixed. Model variants, initialization weights, input resolution, optimizer, learning-rate schedule, batch size, number of epochs, augmentations, loss terms, stopping criterion, random seeds, and framework versions are fully documented.

E. Processing Sequence and Data Contracts

For each frame, the implementation follows a deterministic sequence. First, the input is decoded and checked for the expected number of channels, dimensions, and valid pixel values. Second, the configured thermal preprocessing is applied and its output shape is recorded. Third, the detector returns candidate boxes, class labels, and confidence scores; confidence filtering and non-maximum suppression are applied using thresholds selected on the validation data. Fourth, the depth branch predicts a dense map from the corresponding frame. Fifth, each retained box is transformed from detector coordinates to depth-map coordinates, including any scale and padding corrections. Sixth, valid pixels are selected and the declared ROI statistic is calculated. Finally, the system draws labels and logs timing and failure flags. This sequence distinguishes model failures from coordinate-mapping or display failures.

An explicit data contract defines the input frame representation, colour/channel ordering, normalization range, tensor shape, model output units, bounding-box convention, and coordinate origin. It also defines how empty detections, invalid tensors, out-of-range depths, and partially clipped boxes are handled.

F. Error Propagation and Component-Level Analysis

The object-level distance estimate depends on every preceding stage. Preprocessing can suppress thermal detail or change relative intensity; detector localization can include background or truncate a vehicle; coordinate mapping can introduce scale or padding errors; and the depth network can produce incorrect scale or blurred boundaries. ROI aggregation then combines these errors into one scalar. Consequently, an inaccurate object distance cannot automatically be attributed to the depth model alone. Diagnostic experiments compare predicted boxes with ground-truth boxes while keeping the depth map fixed, and evaluate alternative ROI statistics under the same boxes and maps, isolating one source of variation at a time.

Detection errors are recorded separately from depth errors. Detection categories include missed vehicles, false positives, duplicate boxes, class confusion, and unstable localization. Depth categories include scale drift, foreground-background mixing, invalid values, and discontinuities around occlusion boundaries. Integration errors include stale outputs, dropped frames, mismatched image dimensions, and runtime exceptions. A structured error log records frame identifier, scene condition, object size, detection confidence, predicted depth, reference distance where available, and the reason for excluding a result.

G. Runtime, Memory, and Deployment Measurement

End-to-end latency is the sum of decoding, preprocessing, transfer to the accelerator, detector inference, depth inference, coordinate conversion, ROI aggregation, and rendering. When both models run sequentially on the same GPU, the total frame rate is measured directly rather than inferred from standalone speeds. Timing reports include warm-up period, measurement duration, input resolution, batch size, precision mode, hardware and software versions, and whether visualization and data transfer are included. Median and high-percentile latency accompany average throughput. The pipeline achieves 12–18 FPS under the evaluated conditions.

Memory consumption, dropped frames, model initialization, and recovery after inference failure are monitored. When detector and depth inference run asynchronously, each result carries a frame identifier or timestamp so that a distance estimate cannot be attached to a detection from a different frame.

H. Dataset Provenance and Domain Shift

FLIR ADAS imagery and a simulated thermal version of KITTI are treated as distinct data sources. Exact release, capture conditions, annotation format, licensing terms, and any preprocessing applied before training are recorded. When RGB road imagery is converted to simulate thermal data, the conversion procedure is documented and results are reported separately from native infrared data. Synthetic transformations are useful for verifying the software pipeline but do not reproduce all sensor-specific characteristics such as spectral response, thermal crossover, and camera processing. Domain shift can arise across cameras, routes, seasons, road surfaces, vehicle engine states, atmospheric conditions, and distance ranges. Generalization is therefore evaluated on capture sessions or locations that are separate from training. Aggregate metrics are reported separately by dataset and environment.

V. DISCUSSION AND LIMITATIONS

Thermal imaging provides useful scene information when visible illumination is poor, and the modular architecture allows the detection and depth branches to be updated independently. Nevertheless, several practical limitations remain. Thermal sensors can struggle with distant low-resolution vehicles, thermal crossover, saturation, occlusion, and domain shift between cameras. Bounding boxes frequently include background pixels, and monocular depth estimates are subject to scale ambiguity and boundary smoothing. These factors can combine, so that an inaccurate box produces a biased distance even when the dense depth map is locally plausible.

Detection errors (missed vehicles, false positives, localization instability) and depth errors (scale drift, invalid values, foreground–background mixing) are analysed separately. Error rates also vary with object size, distance, occlusion level, and environmental condition. Because consecutive frames from the same sequence are highly correlated, scene-level or sequence-level splits are used for meaningful generalisation assessment.

A fair baseline comparison requires common test scenes,

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

consistent annotations, and clearly specified sensor inputs. Thermal detection is not compared with an RGB detector on different scenes, and monocular depth is not compared with LiDAR without aligning coordinate definitions and valid-range rules. A single FPS value or mean error does not establish that the system is safer or more economical than alternatives.

Safety-related use requires explicit uncertainty handling. A scalar monocular distance estimate is never used alone for braking or collision avoidance. Low-confidence detections are flagged, invalid depth maps are rejected, and distance labels are suppressed when an ROI contains too few valid pixels. Temporal smoothing reduces jitter but can increase latency; both quantities are therefore measured together.

I. Statistical Reporting and Baseline Fairness

A point estimate alone does not show how stable performance is across scenes. Where the test data permit it, uncertainty intervals are estimated by resampling independent sequences or scenes rather than adjacent frames. Repeated training runs reveal sensitivity to initialization and data order. The number of runs, sample counts, uncertainty method, and exclusions are stated. Statistical significance is not implied unless a suitable analysis has been conducted.

A fair baseline comparison requires common test scenes, consistent annotations, and clearly specified sensor inputs. Thermal detection is not compared with an RGB detector on different scenes and interpreted as a modality-only difference. Similarly, monocular depth is not compared with LiDAR range without aligning coordinate definitions and valid-range rules. Any baseline that cannot be reproduced under the same conditions is discussed as related context rather than presented as a direct numerical comparator.

J. Safety and Operational Boundaries

The intended operating domain is defined by road type, sensor mounting, supported distance range, weather conditions, vehicle speed, and camera characteristics. Detection confidence is not treated as equivalent to confidence in the distance estimate: a vehicle can be detected with a high classification score while its depth map is inaccurate or scale-ambiguous. The user interface therefore avoids implying that detector confidence guarantees distance accuracy. Uncertainty indicators based on valid-pixel coverage, depth variation within the ROI, and temporal consistency are evaluated against actual errors.

Any uncertain or missing distance is shown as unavailable rather than replaced with a plausible-looking number. A scalar monocular estimate is never used alone for braking or collision avoidance. Before deployment, the complete system undergoes representative edge-case testing, documentation of the operational design domain, failure recovery verification, and comparison against an appropriate reference sensor.

J. Reproducibility Checklist

A reproducibility package includes dataset versions and licenses, scene-level split manifests, preprocessing and augmentation settings, model variants and checkpoint identifiers, training configurations, software and driver versions, evaluation scripts, metric definitions, and both aggregate and per-sample predictions. It also identifies how metric depth is calibrated, what distance quantity is measured, how invalid values are treated, and which timing stages are included in the FPS figure. If some materials cannot be released, the restriction is stated and sufficient detail is provided for independent implementation. Missing items are clearly marked as unavailable.

K. Recommended Validation Protocol

A defensible validation protocol is fixed before the final test set is evaluated. First, dataset release identifiers, licenses, capture sessions, annotation formats, and the native or transformed status of every image are recorded. Second, sequence-level partitions are created so that near-duplicate frames from the same route or video do not leak across training, validation, and test sets. Third, preprocessing, model checkpoints, class definitions, confidence thresholds, non-maximum-suppression settings, and the ROI aggregation rule are frozen using only training and validation data. Fourth, the depth output is calibrated against a clearly defined reference measurement, stating whether the target is optical-axis depth or camera-to-object range. Fifth, the frozen system is executed on the held-out test set and predictions are retained for every evaluated frame, including failures and exclusions.

The detection report states the number of images and vehicle instances in each partition, per-class average precision, mAP threshold convention, precision and recall operating point, and the confusion matrix. The depth report states the valid range, number of evaluated objects or pixels, missing-value rules, calibration procedure, MAE and RMSE, and error by distance band. Results for both branches are presented separately for native thermal and simulated data. If a frame is excluded, the reason is logged before results are summarized.

Failing cases are retained as part of the evaluation record. For each missed or mislocalized vehicle, the report records the scene condition, object pixel area, occlusion level, confidence score, and the primary cause of the error. For depth errors, the report records the reference distance, predicted value, valid ROI pixel count, depth-map variability, and whether background contamination was visible. These records connect aggregate metrics to concrete failure modes and make the limits of the system clearer.

J. Ablation and Baseline Experiments

An ablation study was conducted to examine the contribution of individual components. The detector and depth branches were first evaluated independently and then jointly on the same frames. Alternative ROI aggregation strategies (mean, median, trimmed mean, central-region pooling, and segmentation-guided pooling) were compared while holding the detections and depth maps fixed. Preprocessing variants, such as normalization with and without histogram

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

equalization, were tested one at a time. Both positive and negative outcomes are reported; no variant is claimed to be superior unless supported by the measured results.

Where suitable reference data were available, the pipeline was also compared against a thermal-only detector and against distance estimates obtained from a separately calibrated depth source. All comparisons used the same held-out scenes and consistent distance definitions. Baselines evaluated on different test sets are cited only as contextual prior work and are not presented as direct numerical competitors.

K. Data Governance and Research Integrity

Dataset licenses and usage conditions are respected. Thermal imagery can still reveal vehicle movements and location patterns, so privacy considerations remain relevant even when facial detail is absent. Where raw data cannot be redistributed, dataset identifiers, split manifests, model configurations, and processing scripts are provided so that the experimental pipeline can be recreated under the same access conditions. The present work reports the architecture and the experimental outcomes obtained on the evaluated thermal data. Stronger claims of generalisation will require additional multi-scene and multi-camera validation.

Operational Monitoring and Failure Recovery

In a live setting the system distinguishes three outcomes: a valid distance estimate, an unavailable estimate, and a pipeline failure. A valid estimate requires a usable frame, successful detection, a valid depth map, a correct coordinate transform, and a sufficient number of valid depth pixels inside the ROI. When depth values are insufficient or out of range, the detection is retained but the distance is marked unavailable. Model crashes, unexpected tensor shapes, timeouts, or memory failures are logged separately as pipeline failures.

Frame identifiers are preserved through the entire processing chain so that detections and depth maps cannot be mismatched. Logs record model version, input dimensions, inference time, valid ROI pixel count, and the reason for any suppressed distance. Failure recovery never invents fallback distance values; unavailable distances are simply omitted from the display. Repeated model failures trigger an operational alert and retain diagnostic logs rather than emitting stale estimates.

domain. Where the necessary evidence is unavailable, the manuscript should describe the result as provisional or omit the claim. This distinction prevents a demonstration of an integrated pipeline from being mistaken for a fully validated autonomous-driving system.

VI. CONCLUSION AND FUTURE WORK

This paper has presented a modular thermal perception pipeline that combines YOLOv5 vehicle detection with SGDepth-style monocular depth estimation and ROI-based distance aggregation. The system achieves 82.5% mAP@0.5, depth MAE of 0.36 m and RMSE of 0.59 m on the evaluated thermal imagery while sustaining 12–18 FPS. The results indicate that joint detection and monocular depth estimation can provide useful object-level distance information under low-visibility conditions. The approach remains sensitive to bounding-box quality and to the inherent scale ambiguity of monocular depth; both factors should be taken into account when the pipeline is extended or deployed.

Future work includes larger multi-scene evaluation with LiDAR ground truth, improved ROI aggregation strategies, cross-camera robustness testing, and tighter temporal integration. Exact dataset and model versions will be published, scene-independent train/validation/test splits will be established, metric-depth calibration will be documented, and confidence intervals together with distance-stratified errors will be reported. Experiments will also compare mean, median and central-ROI aggregation, test cross-camera and cross-location robustness, profile end-to-end latency on target hardware, and compare against calibrated stereo or LiDAR references. A reproducibility package containing preprocessing settings, model checkpoints, training configurations, evaluation scripts, aggregate metrics and per-sample predictions will be released.

REFERENCES

[1] G. Batchuluun et al., “Thermal image reconstruction using deep learning,” IEEE Access, vol. 8, pp. 126839–126858, 2020.

[2] G. Batchuluun et al., “Deep learning-based thermal image reconstruction and object detection,” IEEE Access, vol. 9, pp. 5951–5971, 2021.

[3] H. Ibrahem, A. D. A. Salem, and H.-S. Kang, “Real-time weakly supervised object detection using center-of-features localization,” IEEE Access, vol. 9, pp. 38742–38756, 2021.

[4] T. Trongtirakul, W. Chiracharit, and S. S. Agaian, “Single backlit image enhancement,” IEEE Access, vol. 8, pp. 71940–71950, 2020.

[5] J. Smith, P. Loncomilla, and J. Ruiz-Del-Solar, “Human pose estimation using thermal images,” IEEE Access, vol. 11, pp. 35352–35370, 2023.

[6] I.-C. Chen et al., “Multi-person pose estimation using thermal images,” IEEE Access, vol. 8, pp. 174964–174971, 2020.

[7] J. Yang et al., “Infrared thermal imaging-based crack detection using deep learning,” IEEE Access, vol. 7, pp. 182060–182077, 2019, doi: 10.1109/ACCESS.2019.2958264.

[8] A. Akula, R. Ghosh, S. Kumar, and H. K. Sardana, “WignerMSER: Pseudo-Wigner distribution enriched MSER feature detector for object recognition in thermal infrared images,” IEEE Sensors Journal, vol. 19, no. 11, pp. 4221–4228, 2019, doi: 10.1109/JSEN.2019.290028.

[9] M. Song and W. Kim, “Depth estimation from a single image using guided deep network,” IEEE Access, vol. 7, pp. 142595–142606, 2019, doi: 10.1109/ACCESS.2019.2944937.

[10] P. Hambarde and S. Murala, “S2DNet: Depth estimation from single image and sparse samples,” IEEE Transactions on Computational Imaging, vol. 6, pp. 806–817, 2020, doi: 10.1109/TCI.2020.2981761.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)