12th INTERNATIONAL SCIENTIFIC CONFERENCE ON DEFENSIVE TECHNOLOGIES
OTEH 2026
Serbia, Tara, 21 - 24 October 2026

IMPACT OF JPEG COMPRESSION ON YOLO-BASED ROAD SCENE OBJECT DETECTION IN VISIBLE AND THERMAL IMAGES

BOBAN PAVLOVIĆ
University of Defence in Belgrade, Military Academy, boban.pavlovic@va.mod.gov.rs

BOBAN BONDŽULIĆ
University of Defence in Belgrade, Military Academy, boban.bondzulic@va.mod.gov.rs

DIMITRIJE BUJAKOVIĆ
University of Defence in Belgrade, Military Academy, dimitrije.bujakovic@va.mod.gov.rs

MOMIR STANKOVIĆ
University of Defence in Belgrade, Military Academy, momir.stankovic@va.mod.gov.rs,

Abstract: *Object detection is a fundamental component of road-scene perception systems used in autonomous driving, where visual data are often compressed to meet storage and bandwidth constraints. While image compression reduces transmission costs, it may introduce distortions that adversely affect detection accuracy. This study investigates the impact of JPEG compression on seven recent YOLO models (YOLOv5m, YOLOv7, YOLOv8m–YOLOv12m), evaluated on visible (VIS) and long wave infrared (LWIR) images from the M^{3}FD dataset. A balanced subset containing daytime and nighttime scenes is used, and ten compressed versions are generated with quality factor (QF) values ranging from 50 to 5. Results show that YOLOv12m achieves the highest global performance for VIS images, particularly during the day, while YOLOv10m excels under low-light conditions. For LWIR data, YOLOv8m performs best overall, with YOLOv12m showing advantages at night. Detection performance decreases gradually for QF values from 50 to 20, but drops sharply below QF=20, highlighting the importance of maintaining moderate image quality. Across intersection over union (IoU) thresholds, LWIR maintains more stable mAP than VIS, demonstrating its robustness under varying illumination. These findings provide quantitative insights into how JPEG compression and imaging modality affect modern YOLO detectors, informing reliable deployment under bandwidth and storage constrained environments.*

Keywords: *road scene object detection, multispectral image, JPEG compression, YOLO.*

1. INTRODUCTION

Autonomous driving systems depend on accurate object detection to perceive complex road scenes, including pedestrians, vehicles, and other traffic participants [1]. These systems rely on visual data for safe navigation and real-time decision making under diverse environmental and lighting conditions. The growing deployment of intelligent transportation systems has vastly increased visual data volumes, straining storage and transmission bandwidth. To mitigate this, image compression techniques such as JPEG are widely used, albeit at the cost of introduced distortions that may affect downstream tasks [2]. Deep learning-based object detection has become a cornerstone of computer vision, enabling machines to locate and classify objects of interest [3]. Because detection is directly tied to image content, compression-induced artifacts can alter edges, textures, and fine details critical for reliable recognition, particularly in adverse conditions like low illumination.

In recent years, YOLO (You Only Look Once) models have achieved notable success across computer vision tasks, including detection, segmentation, and pose estimation [4]. They offer an excellent speed-accuracy trade-off via single-pass inference. YOLO is widely adopted in surveillance, autonomous driving, robotics, medical imaging, and augmented reality [5]. The family has evolved through multiple generations, each addressing prior limitations through architectural innovations, parameter optimization, and improved training methodologies [6], yielding substantial gains in accuracy and efficiency while preserving real-time performance.

Numerous studies have assessed JPEG compression's impact on deep learning-based detection and classification. Dodge and Karam [7] examined effects of Gaussian blur, noise, contrast changes, and JPEG/JPEG2000 compression on classification models, finding CNNs particularly sensitive to blur and noise, with notable accuracy degradation. In [8], authors showed that compression artifacts degrade CNN-based object detectors' accuracy. Similarly, [2] analyzed JPEG compression on RetinaNet and Faster R-CNN, reporting pronounced detection drops at high compression levels.

Bayerl et al. [9] studied remote sensing image compression (JPEG, JPEG2000) on classification, showing that mismatched compression settings between

Corresponding author: B. Pavlović;
E-mail: boban.pavlovic@va.mod.gov.rs
DOI: 10.5937/OTEH26060P

training and inference cause substantial accuracy loss. In [10], authors evaluated HEVC compression on object tracking using YOLOv3 and SORT, finding that quantization significantly affects tracking performance while motion search range does not, and derived a regression model linking compression strength to accuracy. In [11], the robustness of several YOLO detectors to common distortions in remote sensing was evaluated, including Gaussian noise, JPEG/JPEG2000 compression, and blur. All distortions degraded accuracy, with YOLOv9 most robust and YOLOv7 and YOLOv10 most sensitive.

Despite these efforts, the robustness of YOLO detectors on infrared imagery, especially compared to visible (VIS) images under compression, remains insufficiently explored. Most existing studies focus on RGB data, leaving JPEG compression effects on long-wave infrared (LWIR) largely unexamined. Moreover, comparative VIS–LWIR analyses are limited, despite their distinct statistical and perceptual properties, which may lead to different responses to compression artifacts affecting textures and object boundaries. To fill this gap, this paper analyzes seven YOLO models – YOLOv5m [12], YOLOv7 [13], YOLOv8m [4], YOLOv9m [14], YOLOv10m [15], YOLOv11m [6], and YOLOv12m [16] – using a balanced test set from the M^3FD dataset, containing equal VIS and LWIR images split evenly between daytime and nighttime conditions. Both modalities are compressed with JPEG quality parameters from 50 down to 5 to systematically evaluate compression impact on detection performance.

The remainder of this paper is structured as follows. Section 2 describes the methodology, covering the dataset, YOLO models, and JPEG compression process. Section 3 presents experimental results with analysis and discussion. Section 4 summarizes key findings and outlines future research directions.

2. METHODOLOGY

2.1. Dataset description

In this study, VIS and LWIR images were used from the M^3FD [17] dataset, which is widely adopted for evaluating object detection performance in driving scenarios. A total of 440 VIS images and 440 LWIR images were selected to represent diverse driving environments. For each modality, 220 images correspond to daytime scenes and 220 images to nighttime conditions, ensuring a balanced evaluation across lighting scenarios. The selected images include typical road scenes containing pedestrians (referred to as person), cars, and buses, which are the three object categories considered in this study. The selected set covers both urban and suburban environments and includes variations in object scale, occlusion, and background complexity, providing a challenging and realistic benchmark for assessing detection performance under compression.

Table 1. Distribution of object instances in the selected subset across daytime, nighttime, and combined scenarios

Class	Daytime vision	Nighttime vision	All-day/night vision
Person	117	206	323
Car	219	220	439
Bus	147	10	157
All	**483**	**436**	**919**

Figure 1 shows images from both modalities - visible and infrared based on both daytime and nighttime conditions. The selected examples illustrate how a single image can contain representatives of multiple classes, as well as multiple samples within the same class. Additionally, it can be observed that objects may be close or far away, which would test the robustness of the detectors themselves. Objects may be visible in one modality but not visible in the other (see the presence of a person in nighttime conditions in LWIR, and their absence in VIS).

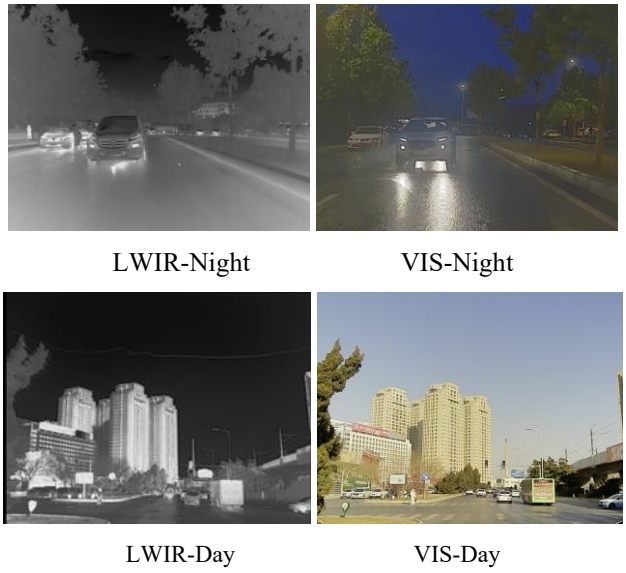

Figure 1. Example images from visible and infrared modalities based on daytime and nighttime conditions

2.2. YOLO models

Seven YOLO models are selected in this study (YOLOv5, YOLOv7–YOLOv12) to provide a comprehensive comparison across different design stages and performance trade-offs. YOLOv5m [12] introduced a stable anchor-based design using CSPDarknet53 with FPN and PANet for multi-scale features extraction. YOLOv7 [13] enhanced this design with E-ELAN and improved scaling strategies to better balance accuracy and speed. YOLOv8m [4] adopted an anchor-free framework, incorporating C2f modules and a decoupled head to improve gradient flow and detection precision. YOLOv9m [14] further improved training stability through programmable gradient information and employed the GELAN backbone for higher efficiency. YOLOv10m [15] focused on real-time performance by introducing a fully NMS-free pipeline with dual label assignment to reduce latency while improving accuracy. YOLOv11m [6] extended these developments with advanced attention mechanisms, such as C2PSA, C3K2 modules, and the SPPF block to better capture spatial

context while maintaining fast inference. The latest version, YOLOv12m [16], continues this evolution by refining architectural efficiency and training strategies to improve accuracy, robustness, and deployment flexibility.

2.3. JPEG compression

JPEG is a widely used lossy image compression method that reduces image size while maintaining acceptable visual quality [18]. It works by transforming the image into the frequency domain using the Discrete Cosine Transform (DCT), which separates visual information into low- and high-frequency components. Low frequencies capture most of the image structure, while high frequencies represent fine details and noise. The resulting DCT coefficients are then quantized to remove perceptually less important information, achieving significant data reduction. JPEG is valued for its strong balance between compression efficiency, visual fidelity, and low computational cost, enabling fast encoding and decoding. The compression level is controlled by a quality factor (QF) ranging from 1 to 100, where higher values correspond to lower compression and better image quality. In this study, JPEG compression was applied at 10 levels by varying the QF from 5 to 50 in steps of 5 using the MATLAB *imwrite* function.

3. RESULTS AND DISCUSSION

3.1 Experimental setup

The experimental setup was conducted on a Windows 11 operating system, running on an Intel® Core™ i7-12650H processor (2.30 GHz) with 16 GB of RAM, and an NVIDIA RTX 3050Ti GPU with 4 GB of memory. All algorithms were implemented using the PyTorch framework (version 2.0.1) and accelerated with the CUDA Toolkit (version 11.8).

In this study, all YOLO models are evaluated using pretrained weights from the MS-COCO dataset [19], which contains 80 object classes, including those selected from the M^3FD dataset. Each model is tested on the original image subsets as well as on ten JPEG-compressed versions generated by varying the QF. The evaluations are conducted separately for VIS and LWIR modalities and further divided into daytime and nighttime conditions to analyze the influence of illumination, providing a comprehensive comparison of detection performance across compression levels, imaging modalities, and lighting scenarios.

To assess the robustness of the YOLO models, widely adopted evaluation metrics were employed, including precision (P), recall (R), and mean average precision (mAP). The mAP was computed across intersection over union (IoU) thresholds ranging from 0.5 to 0.95 in increments of 0.05, following the MS-COCO evaluation protocol. Additionally, the mean average precision at IoU = 0.5 ($mAP_{0.5}$), consistent with the PASCAL VOC metric, was reported [20].

3.2 Global detection performance

Table 2 presents the quantitative comparison of YOLOv5, YOLOv7–YOLOv12 on the VIS-Day, VIS-Night, and both VIS subsets in terms of precision (P), recall (R), $mAP_{0.5}$, and mAP averaged across the ten JPEG compression levels from 50 to 5 with step of 5. The best result for each metric within a given subset is highlighted in **bold**.

For the VIS-Day subset, YOLOv12m achieves the highest precision (0.8055) and mAP (0.4625), indicating superior detection accuracy under daytime conditions. YOLOv7 attains the best recall (0.6462), while YOLOv7 and YOLOv12m provide the strongest $mAP_{0.5}$ performance, with YOLOv7 slightly higher (0.7073). Overall, YOLOv12m demonstrates the most balanced and robust performance in this subset.

In the VIS-Night subset, performance degradation is observed across all models due to challenging illumination conditions. YOLOv5m and YOLOv12m obtain the highest precision (0.4375 and 0.4372, respectively), while YOLOv10m achieves the best recall (0.2325), $mAP_{0.5}$ (0.26), and mAP (0.1517), highlighting its improved robustness in low-light scenarios.

Table 2. Detection results across daytime, nighttime, and both scenarios for VIS test set

Subset	Metric	YOLOv5	YOLOv7	YOLOv8	YOLOv9
VIS-Day	P	0.7221	0.7720	0.7657	0.7221
	R	0.6283	**0.6462**	0.6268	0.6305
	$mAP_{0.5}$	0.6736	**0.7073**	0.6905	0.6870
	mAP	0.4286	0.4496	0.4533	0.4458
VIS-Night	P	**0.4375**	0.3651	0.3980	0.4243
	R	0.2001	0.2052	0.2174	0.1823
	$mAP_{0.5}$	0.2438	0.2247	0.2492	0.2257
	mAP	0.1342	0.1235	0.1448	0.1310
VIS	P	0.7110	0.7068	0.7319	0.6646
	R	0.5276	**0.5332**	0.5285	0.5103
	$mAP_{0.5}$	0.5774	0.5808	0.5882	0.5573
	mAP	0.3705	0.3784	0.3918	0.3691

Subset	Metric	YOLOv10	YOLOv11	YOLOv12
VIS-Day	P	0.7687	0.7445	**0.8055**
	R	0.6207	0.6304	0.6349
	$mAP_{0.5}$	0.6853	0.6847	0.7048
	mAP	0.4480	0.4438	**0.4625**
VIS-Night	P	0.3891	0.3555	0.4372
	R	**0.2325**	0.1899	0.2070
	$mAP_{0.5}$	**0.2600**	0.2139	0.2462
	mAP	**0.1517**	0.1233	0.1425
VIS	P	0.7227	0.7149	**0.7754**
	R	0.5286	0.5027	0.5233
	$mAP_{0.5}$	0.5822	0.5590	**0.5908**
	mAP	0.3878	0.3699	**0.3951**

For the complete VIS set, YOLOv12m consistently outperforms other models, achieving the highest precision (0.7754), $mAP_{0.5}$ (0.5908), and mAP (0.3951). The highest recall is obtained by YOLOv7 (0.5332), though YOLOv12m remains competitive.

These results confirm that recent YOLO versions, particularly YOLOv12m and YOLOv10m, benefit from architectural and training improvements, delivering consistent performance gains over earlier versions, especially in terms of mean average precision, while maintaining competitive recall and precision across varying illumination conditions.

Table 3 presents performance comparison of YOLOv5m, YOLOv7, YOLOv8m–YOLOv12m on the LWIR-Day, LWIR-Night, and both LWIR subsets in terms of precision, recall, $mAP_{0.5}$, and mAP, averaged over ten JPEG compression levels.

Table 3. Detection results across daytime, nighttime, and both scenarios for LWIR test set

Subset	Metrics	YOLOv5	YOLOv7	YOLOv8	YOLOv9
LWIR-Day	P	0.4452	0.4424	**0.5153**	0.4398
	R	0.3951	0.4213	**0.4229**	0.3193
	$mAP_{0.5}$	0.3950	0.4213	**0.4460**	0.3603
	mAP	0.2271	0.2443	**0.2612**	0.2038
LWIR -Night	P	**0.4641**	0.4476	0.4260	0.4201
	R	0.3780	0.3815	0.4133	0.3797
	$mAP_{0.5}$	0.4009	0.3984	0.4222	0.4030
	mAP	0.2349	0.2347	0.2624	0.2495
LWIR	P	0.5243	0.5333	**0.5853**	0.4973
	R	0.4743	0.4814	**0.5057**	0.4050
	$mAP_{0.5}$	0.4887	0.5015	**0.5351**	0.4458
	mAP	0.2934	0.3062	**0.3370**	0.2726

Subset	Metrics	YOLOv10	YOLOv11	YOLOv12
LWIR-Day	P	0.4794	0.4695	0.4634
	R	0.3209	0.3303	0.3035
	$mAP_{0.5}$	0.3913	0.3683	0.3524
	mAP	0.2232	0.2043	0.1948
LWIR -Night	P	0.4407	0.4623	0.4637
	R	0.3911	0.4012	**0.4245**
	$mAP_{0.5}$	0.4239	0.4182	**0.4383**
	mAP	0.2631	0.2552	**0.2726**
LWIR	P	0.5219	0.5295	0.5384
	R	0.4247	0.4154	0.4108
	$mAP_{0.5}$	0.4897	0.4522	0.4657
	mAP	0.3014	0.2709	0.2838

For the LWIR-Day subset, YOLOv8m clearly outperforms the other models, achieving the highest precision (0.5153), recall (0.4229), $mAP_{0.5}$ (0.4460), and mAP (0.2612). This indicates that YOLOv8m provides the most accurate and reliable detections under daytime conditions.

In the LWIR-Night subset, YOLOv12m attains the best overall performance, with the highest recall (0.4245), $mAP_{0.5}$ (0.4383), and mAP (0.2726). The highest precision is obtained by YOLOv5m (0.4641), closely followed by YOLOv12m (0.4637). These results show that YOLOv12m offer improved robustness in nighttime LWIR scenarios, where thermal contrast is more pronounced.

For the complete LWIR set, YOLOv8m consistently achieves the best results across all metrics, including precision (0.5853), recall (0.5057), $mAP_{0.5}$ (0.5351), and mAP (0.3370). This confirms its strong generalization capability across both day and night thermal conditions.

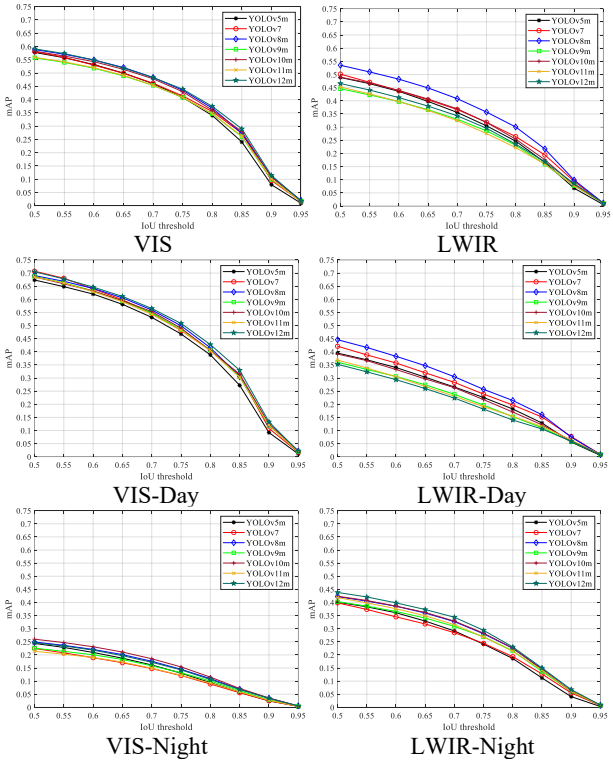

Figure 2. mAP as a function of the IoU threshold (0.5–0.95) for VIS and LWIR modalities under day and night conditions, averaged over ten JPEG compression levels (QF = 50 to 5)

The obtained results demonstrate that YOLOv8 delivers the most consistent and superior performance on LWIR data, particularly in the complete and daytime subsets, while YOLOv12m shows notable advantages under nighttime conditions.

3.3. Detection performance across IoU thresholds under JPEG compression

Figure 2 illustrates the variation in detection performance, measured in terms of mAP, as a function of the IoU, which ranges from 0.5 to 0.95 in steps of 0.05, for both visible and LWIR modalities under daytime and nighttime conditions. For each IoU threshold, the mAP is computed and then averaged over ten JPEG compression levels.

Across all configurations, the mAP decreases monotonically as the IoU threshold increases, reflecting the stricter localization requirement imposed by higher IoU values. For the VIS modality, daytime scenes consistently achieve higher mAP values than nighttime scenes at all IoU thresholds, while the performance gap becomes more pronounced at higher thresholds, indicating reduced localization accuracy under low-light conditions.

For the LWIR modality, there is a smaller gap between mAP curves in day and night scenarios. This behavior suggests that thermal imagery provides more consistent object localization under varying illumination conditions. Nevertheless, a clear performance drop is still observed.

3.4 Detection performance under varying JPEG compression levels

Figure 3 illustrates the mAP results of different YOLO object detection models under JPEG compression across QF values from 50 to 5 with step of 5, for both VIS and LWIR modalities under overall, daytime, and nighttime conditions.

Across all scenarios and YOLO models, a consistent trend is observed regarding image compression. From QF=50 to QF=20, the models remain relatively stable, demonstrating robustness to moderate compression. The mAP shows a gradual decline with a gentle slope, indicating only minor performance degradation. However, when the quality factor drops below 15, the curves exhibit a sharp "knee," with detection performance decreasing steeply as QF approaches 5. This indicates that maintaining a quality factor above 20 is essential to preserve reliable detection under compressed conditions.

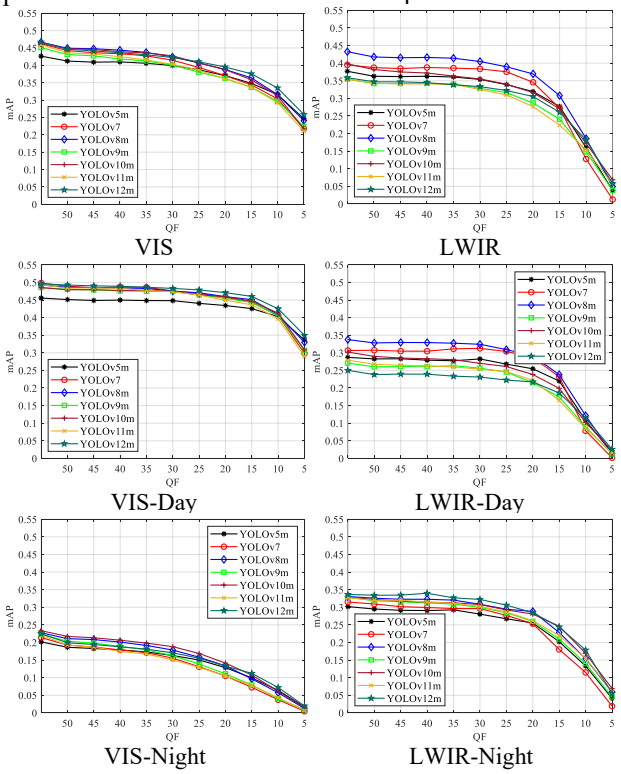

Figure 3. mAP results of YOLO models under JPEG compression across QF values for VIS and LWIR images under different lighting conditions

In comparing the models, clear differences in performance emerge. YOLOv12m and YOLOv8m consistently outperform the others across most modalities, starting from higher baselines at QF=50 and maintaining their lead as image quality decreases. The exception is VIS under nighttime conditions, where YOLOv10m achieves superior mAP values at QF levels above 20. The remaining YOLO models occupy the mid-range and frequently exchange positions. They generally outperform YOLOv5m for the visible modality and LWIR under nighttime conditions, but lag behind YOLOv8m and YOLOv12m in overall performance.

Analyzing performance across modalities reveals clear differences between VIS and LWIR images. During daytime, VIS cameras dominate, achieving mAPs near 0.50, while LWIR-Day lags behind at 0.30–0.35. This confirms that visual data provides richer texture and detail under sufficient illumination. At night, the situation reverses. VIS-Night performance drops sharply, with even the best models reaching only about 0.23 mAP, illustrating the limitations of standard cameras in low light environments. In contrast, LWIR-Night, maintains robust performance, peaking around 0.34 mAP, exceeding VIS-Night and approaching to LWIR-Day scenario. These findings indicate that thermal imaging offers consistent detection across day and night, whereas visible-light cameras are highly sensitive to lighting conditions.

3.5 Class-wise detection performance under JPEG compression

Figure 4 presents a detailed class-wise analysis of detection performance for the person, car, and bus categories. The mAP is averaged over ten JPEG compression levels.

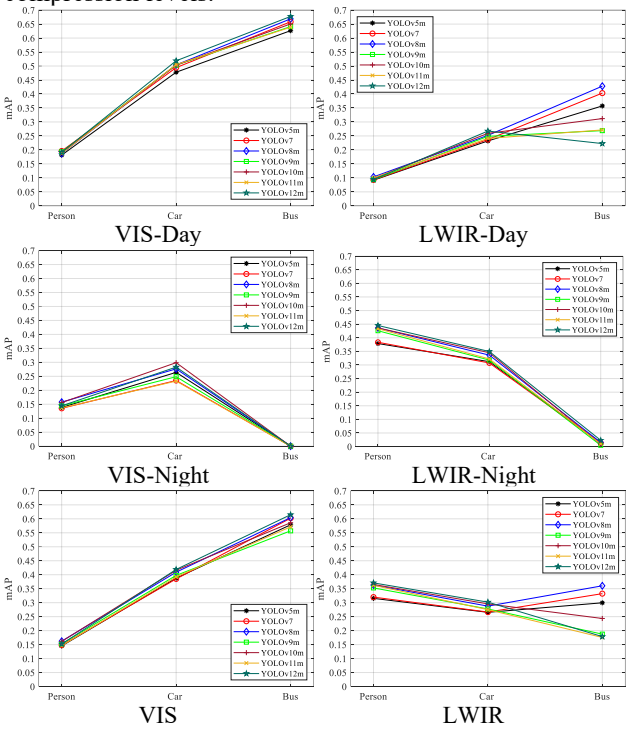

Figure 4. Class-wise detection performance in term of mAP for person, car, and bus, averaged over ten JPEG compression levels, for VIS and LWIR under daytime and nighttime conditions

In the VIS modality, daytime conditions outperform nighttime conditions for all classes, with the performance gap being most pronounced for the bus class. This indicates that YOLO detectors rely heavily on texture and color that are significantly affected by both low-light conditions and JPEG compression. The global VIS results follow the same trend as daytime conditions, confirming the dominant influence of illumination on visible-spectrum detection.

For LWIR modality, nighttime demonstrates the highest robustness for class person. The performance difference between LWIR-Day and LWIR-Night is smaller than in VIS for class car, highlighting the reduced sensitivity of thermal imagery to illumination changes.

Across all scenarios, the car class consistently achieves a stable mAP, reflecting its stable visual and thermal signatures, which make it less sensitive to compression artifacts. The person class exhibits moderate performance, with noticeable differences between day and night conditions, particularly in the LWIR modality, where reduced illumination and compression jointly degrade fine-grained features. The bus class shows more variable behavior, which can be attributed to its lower number of instances within nighttime scenario.

4. CONCLUSION

This study provides a comprehensive evaluation of JPEG compression effects on road scene object detection using YOLO models from YOLOv5m to YOLOv12m across visible and LWIR imagery. The results demonstrate that recent YOLO architectures, particularly YOLOv12m and YOLOv8m, benefit from architectural and training improvements, achieving higher precision, recall, and mAP under both daytime and nighttime conditions. VIS detection performance is highly sensitive to illumination and compression, with daytime scenes outperforming nighttime, whereas LWIR offers more consistent detection across different lighting conditions. Compression impacts all models, with a gradual mAP decline for QF values above 20 and a steep drop below QF=20, emphasizing the need to maintain moderate image quality. The analysis also reveals modality-specific trends, VIS provides richer texture under sufficient light, while LWIR ensures robust detection at night. These findings underscore the importance of considering both compression level and imaging modality when deploying real time road scene object detection systems, and they provide practical guidelines for maintaining reliable performance in bandwidth limited scenarios. Future work may explore alternative compression schemes, additional detection architectures, and adaptive training strategies on compressed data to further improve detection performance.

References

[1] ADAM M.A.A., TAPAMO, J.R.: *Survey on image-based vehicle detection methods,* World Electric Vehicle Journal 16(6), 303 (2025).

[2] GANDOR, T., NALEPA, J.: First gradually, then suddenly: understanding the impact of image compression on object detection using deep learning, Sensors 22(3), 1104 (2022).

[3] BAKIR, H., Bakir, R.: Evaluating the robustness of YOLO object detection algorithm in terms of detecting objects in noisy environment, Journal of Scientific Reports-A 054, 1–25 (2023).

[4] JOCHER, G., AYUSH, C., JING, Q.: *Ultralytics YOLOv8,* (2023). [Online]. Available: https://github.com/ ultralytics/ ultralytics. Accessed 2 Aug 2025

[5] Terven, J., CÓRDOVA-ESPARZA, D.-M., ROMERO-GONZÁLEZ, J.-A.: *A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS,* Machine Learning and Knowledge Extraction 5(4), 1680–1716 (2023).

[6] Glenn, J., JING, Q.: *Ultralytics YOLOv11,* (2024). [Online]. Available: https://github.com/ultralytics/ ultralytics. Accessed 2 Aug 2025

[7] DODGE, S., KARAM, L.: *Understanding how image quality affects deep neural networks,* In: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1–6, Lisbon, Portugal, (2016).

[8] AQQA, M., MANTINI, P., SHAH, S.K.: *Understanding how video quality affects object detection algorithms,* In: Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Application, pp. 96–104, Prague, Czech Republic, (2019).

[9] BAYERL, A., KEGLEVIC, M., WÖDLINGER, M.G., SABLATNIG, R.: *Impact of learned domain specific compression on satellite image object classification*, In: Proceedings of the 26th Computer Vision Winter Workshop (CVWW 2023); Austria, (2023).

[10] Tanaka, T., HAREL, A., BAJIĆ, I.V.: *Does video compression impact tracking accuracy?,* In: 2022 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1517–1521, Austin, TX, USA, (2022).

[11] ADLI, T., BUJAKOVIĆ, D.M., BONDŽULIĆ, B.P., Laidouni, M.Z., ANDRIĆ, M.S.: *Robustness of YOLO models for object detection in remote sensing images,* Journal of Electrical Engineering 76(5), 429–442 (2025).

[12] Jocher, G.: *YOLOv5 by Ultralytics,* (2020). [Online]. Available: https://github.com/ultralytics/yolov5. Accessed 2 Aug 2025

[13] WANG, C.Y., BOCHKOVSKIY, A., LIAO, H.Y.M.: *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, pp. 7464–7475, Vancouver, BC, Canada, (2023).

[14] WANG, C.Y., YEH, I.H., LIAO, H.Y.M.: YOLOv9: *Learning what you want to learn using programmable gradient information,* In: European Conference on Computer Vision, pp. 1–21, Milan,

Italy, (2024).

[15] CHEN, H., et al.: YOLOv10: *Real-time end-to-end object detection,* In: Advances in Neural Information Processing Systems 37 (NeurIPS 2024), pp. 107984–108011, San Diego, California, USA, (2024).

[16] Tian, Y., YE, Q., DOERMANN, D.: YOLOv12: *Attention-centric real-time object detectors,* arXiv (2025). arXiv:2502.12524.

[17] LIU, J., FAN, X., HUANG, Z., WU, G., LIU, R., ZHONG, W., LUO, Z.: Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection, In: Proceedings of the IEEE/CVF Conference on Com-puter Vision and Pattern Recognition, pp. 5802–5811, New Orleans, LA, USA, (2022).

[18] Hudson, G., LÉGER, A., NISS, B., SEBESTYÉN, I., VAABEN, J.: *JPEG-1 standard 25 years: past, present, and future reasons for a success,* Journal of Electronic Imaging 27(4), 040901 (2018).

[19] LIN, T.-Y., MAIRE, M., BELONGIE, S., et al.: *Microsoft COCO: Common objects in con-text,* In: European Conference on Computer Vision, pp. 740–755, Zurich, Switzerland (2014).

[20] Padilla, R., NETTO, S.L., DA SILVA, E.A.B.: *A survey on performance metrics for object-detection algorithms,* In: 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 237–242, Niteroi, Brazil, (2020).