applied sciences

MDPI

Review

Visible–Infrared Image Fusion for Computer Vision: A Review of Datasets and Fusion Strategies in Object Detection and Facial-Expression Recognition

Muhammad Tahir Naseem [1], Chan-Su Lee [1,*] and Muhammad Adnan Khan [2,*]

[1] Department of Electronic Engineering, Yeungnam University, Gyeongsan-si 38541, Republic of Korea; nmtahir@yu.ac.kr
[2] Department of Software, Faculty of Artificial Intelligence and Software, Gachon University, Seongnam-si 13120, Republic of Korea
* Correspondence: chansu@ynu.ac.kr (C.-S.L.); adnan@gachon.ac.kr (M.A.K.)

Abstract

Visible and infrared (IR) image fusion has become an important strategy for improving computer vision performance under low illumination, occlusion, and some poor-visibility conditions. By integrating complementary textural information from visible images with thermal or IR cues, VIR fusion can enhance object localization, detection robustness, and facial-expression recognition (FER). This review examines VIR fusion techniques and datasets for computer vision applications, with object detection (OD) considered as a relatively mature scene-level task and FER considered as an emerging human-centered application. It summarizes major multimodal datasets, compares early-fusion approaches, including sensor- and feature-level fusion, with late-fusion approaches, including score- and decision-level fusion, and discusses representative machine learning and deep learning methods. The review also evaluates commonly used performance metrics and identifies current limitations, including dataset imbalance, sensor misalignment, limited demographic diversity in facial-expression datasets, computational complexity, and weak real-time generalization. Finally, key application areas, including surveillance, healthcare, remote sensing, autonomous systems, and human–computer interaction, are discussed. This review highlights the need for better-aligned multimodal datasets, standardized evaluation protocols, lightweight fusion architectures, and robust models capable of operating in dynamic real-world environments.

Keywords: visible–infrared fusion; thermal imaging; object detection; facial-expression recognition; multimodal learning; deep learning; image-fusion datasets; sensor fusion

Academic Editors: Shangshu Yu, Peter Peer and Tsung-Jung Liu

Received: 22 May 2026
Revised: 19 June 2026
Accepted: 22 June 2026
Published: 6 July 2026

1. Introduction

Image fusion involves applying specific computer algorithms to combine heterogeneous images, typically captured by different sensor types, into a single, fused image that incorporates data from multiple sources. The obtained fused image is more readily perceived by the human visual system and is better suited for other high-level vision tasks, such as semantic segmentation, object detection (OD), and tracking [1–3]. Infrared (IR) images are generated from thermal radiation emitted by objects and are therefore useful for highlighting heat-emitting targets, particularly in darkness or low-light scenes. However, IR images often lack fine texture, color, and surface-detail information, and their quality can be degraded by fog, rain, glass, atmospheric absorption, low thermal contrast, sensor noise,

and thermal reflections. In contrast, visible images are mainly formed from reflected visible light and provide richer texture, edge, color, and appearance information. However, visible images are strongly affected by illumination changes, shadows, smoke, haze, and nighttime conditions. Therefore, VIR fusion aims to combine the thermal-saliency information from IR imagery with the texture and structural details from visible imagery, thereby leveraging the complementary strengths and limitations of the two modalities.

Fused images can better capture the scene information of a target than single-source images, thus significantly improving image clarity and quality. The efficacy of a fusion algorithm is determined by its capability to efficiently extract features and detailed information from a source image without introducing additional noise during fusion. The two general categories of multisensory fusion architectures, early and late fusion, are determined by the degree of data abstraction employed during fusion [4]. In the early fusion architecture, raw data from two sensors are combined at the pixel level, after which features from both sensors are integrated at the feature level. Conversely, the late-fusion architecture comprises decision-level fusion, which facilitates separate detection from each sensor, and score-level fusion, which integrates the obtained feature scores. Notably, final detection is achieved through subsequent decision-level integration of sensor outputs.

Autonomous vehicles encounter critical object-recognition issues; studies have extensively attempted to improve their object-recognition speed and efficiency. Accurate OD is essential to a robust perception of the surroundings offered by a dependable autonomous driving system. Additionally, OD is critical to the success of subsequent tasks, such as object classification and tracking. The dynamic nature of the scene, weather, view distances, and fluctuating light levels complicate OD in aquatic environments. Moreover, false detections may be due to changes in lighting, camera motion, and light reflection [5]. Convolutional neural networks (CNNs) have been used to significantly improve the performance of computer vision tasks, such as object classification [6], detection [7,8], and segmentation [9]. Furthermore, CNNs have been applied to autonomous vehicles using a variety of fusion methodologies [5,10,11], with some focusing on RGB images and others incorporating IR images for object recognition.

A previous study [12] employed the classic YOLOv5 model to demonstrate a multimodal fusion technique for OD. The authors developed an image-target detection model using YOLOv5mb, which was specifically designed for aerial platforms. It excellently synthesizes missile-captured images. Further, this model accurately identifies targets in single-mode and multimodal VIR missile-captured images. Additionally, the fusion-layer architecture of YOLOv5mb makes it very suitable for object recognition in missile-captured multimodal VIR images. A previous study [13] demonstrated two deep learning (DL) models and various fusion techniques for animal detection and classification, revealing the suitability of image fusion for surreptitious animal monitoring in their natural habitats. A previous study [14] presented a unique VIR image-fusion technique based on the Gaussian estimation (GE)-weighted average (GE-WA) model and the modified VGG-19 architecture. Further, the visible and IR images in this model were first divided into basic images and detailed contents using the Laplacian.

Furthermore, to produce a fundamental fusion image that eliminates the halo effect in a visible image, a GE function was developed, and a basic fusion method was subsequently implemented based on the GE-WA model. Thereafter, various depth features were extracted from the visible and IR images using the multilayer fusion approach and the pre-trained VGG-19 network, thereby enabling the extraction of fused detail information. After the fusion, the basic picture and detailed content were used to recreate the fused image.

The FER component of this review is motivated by multimodal visible–IR facial-expression datasets such as the VIRI dataset introduced by Siddiqui et al. [15]. VIRI

provides paired visible and IR facial-expression images for multimodal emotion-recognition research and demonstrates the relevance of VIR information for human-centered affective computing. Accordingly, Figure 1 presents a task-oriented schematic of VIR fusion for both OD and FER, illustrating how visible and IR inputs can be integrated at pixel, feature, score, and decision levels for downstream computer vision tasks.

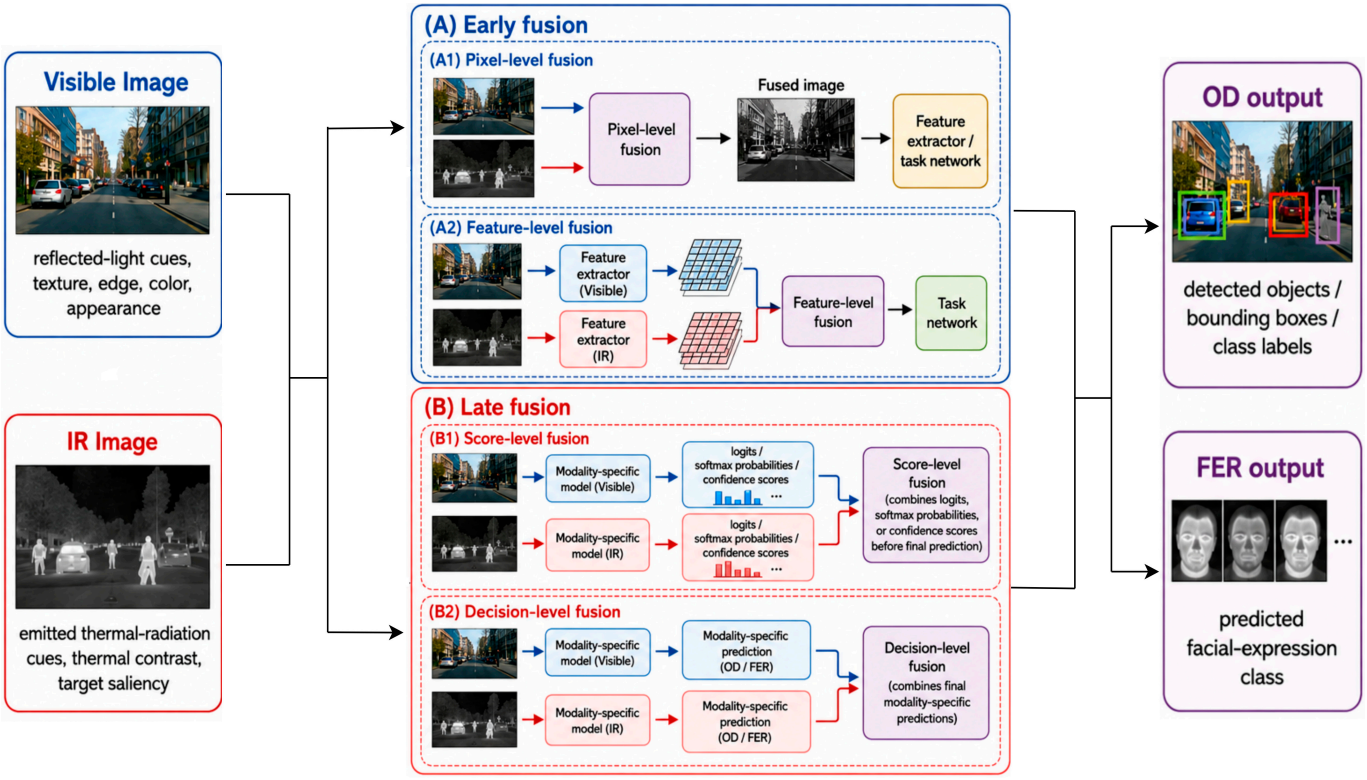

Figure 1. Task-oriented overview of VIR fusion for computer vision applications, with emphasis on OD and FER. Visible images provide reflected-light texture, edge, color, and appearance cues, whereas IR images provide emitted thermal-radiation cues, thermal contrast, and target saliency. Visible and IR information can be integrated through early fusion, including pixel-level and feature-level fusion, or late fusion, including score-level and decision-level fusion, to support downstream OD and FER tasks. In the figure, blue arrows indicate the visible-image stream, red arrows indicate the infrared-image stream, and black arrows indicate the fused/task-level processing flow; the colored boxes distinguish inputs, fusion modules, task networks, and output blocks, while the ellipsis indicates additional representative FER outputs or classes.

FER is a key area of human–computer interaction that aims to interpret emotional states from facial cues using computational models [16,17]. FER is a pivotal aspect of creating empathetic and responsive computer systems that have traditionally relied on a single modality, such as FEs or vocal intonation [18,19]. However, integrating several modalities, such as physiological signals and contextual data, has been proven to significantly improve FER precision and resilience. However, this multimodal approach poses unique methodological challenges, ranging from the synchronization and fusion of disparate data types to the interpretation of complex, nuanced emotional states [20]. Multimodal FER proceeds via a wide variety of techniques—from feature-level fusion (where different data types are integrated before analysis) to decision-level fusion (where independent decisions from multiple models are integrated to produce a final verdict on the subject's emotional state) [20,21]. The variability of individual emotional responses, even within the same contextual parameters, further increases the complexity of creating accurate and reliable recognition systems [22,23]. Notwithstanding these obstacles, advancements in data fusion

and machine learning (ML) methodologies continue to expand the potential applications of multimodal FER, creating new research and implementation platforms across industries such as customer service, education, and mental health monitoring [20,24]. In this review, these HCI, speech, EEG, and physiological-signal studies are cited only to motivate the broader multimodal FER context; the core methodological discussion is restricted to VIR image-based datasets, fusion strategies, and downstream OD and FER applications. Figure 1 illustrates a task-specific overview of the OD and FER framework based on VIR image fusion. The process begins by localizing relevant regions, such as objects, heads, or faces, in visible and IR image pairs. These localized regions are then processed using deep learning techniques for downstream tasks such as OD and FER. The diagram summarizes two broad fusion strategies, early and late fusion, used to integrate complementary information from visible and IR images. Early fusion includes pixel-level and feature-level integration before or during feature extraction, whereas late fusion includes score-level integration of logits, softmax probabilities, or confidence scores and decision-level integration of final detections or expression predictions. This task-specific description clarifies the pixel-, feature-, score-, and decision-level dataflow used in VIR-based OD and FER.

In this study, we review multimodal VIR datasets and fusion techniques, highlighting their role in enhancing computer vision performance under challenging imaging conditions. The scope of this review is VIR image fusion, rather than OD or FER as independent research topics. OD and FER are discussed as two representative downstream computer vision applications that benefit from complementary visible and IR information. OD represents scene-level perception tasks, where VIR fusion improves target localization and recognition under low illumination, occlusion, adverse weather, and complex background conditions. FER represents human-centered affective analysis, where visible imagery provides texture and appearance cues, while IR imagery provides complementary illumination-invariant and thermal information. Unlike OD, which has a comparatively larger body of RGB-T and VIR studies, VIR-based FER is still an emerging research area; therefore, this review uses FER to highlight the opportunities and limitations of applying VIR fusion to human-centered affective computing.

Recent surveys have reviewed infrared–visible image-fusion algorithms, neural network-based fusion methods, and broad VIR image-fusion applications [25,26]. However, these surveys mainly emphasize general fusion algorithms, image-quality improvement, or broad application domains, and they do not specifically synthesize VIR fusion from the joint perspective of OD and FER. Therefore, the novelty of the present review lies not in claiming the absence of VIR fusion surveys but in providing a task-oriented synthesis that connects multimodal datasets, fusion-level design, task-specific evaluation metrics, OD-oriented studies, FER-oriented studies, cross-study comparison, applications, and future challenges within one computer vision framework. This organization makes it possible to compare a relatively mature scene-level task, OD, with an emerging human-centered task, FER, and to identify both shared and task-specific limitations.

Accordingly, this review aims to clarify how visible and IR information is integrated at different fusion levels, how these fusion strategies affect downstream OD and FER performance, and why evaluation protocols must be interpreted according to task type. The review covers early fusion, including sensor-/pixel-level and feature-level fusion, and late fusion, including score-level and decision-level fusion. In addition, VIR and related RGB-T datasets, representative application-specific studies, evaluation metrics, current limitations, deployment challenges, and future research opportunities are discussed.

Review Methodology

To improve the transparency and reproducibility of this review, a structured literature-search and study-selection procedure was followed. Relevant studies were identified from major literature sources, including Web of Science, Scopus, IEEE Xplore, ACM, ScienceDirect, SpringerLink, and Google Scholar. The search used combinations of keywords related to "visible–infrared image fusion," "infrared–visible image fusion," "thermal-visible fusion," "RGB-T fusion," "multimodal object detection," "visible–thermal object detection," "facial-expression recognition," "thermal facial-expression recognition," and "multimodal emotion recognition." The retrieved studies were first screened based on their titles, abstracts, keywords, and relevance to the review scope. Full-text eligibility assessment was then performed to determine whether each study addressed VIR or thermal–visible fusion, multimodal datasets, fusion strategies, evaluation metrics, object detection, FER, or application-specific analysis. Studies were included if they provided relevant technical details, dataset descriptions, fusion methods, experimental results, or critical discussion related to visible–infrared fusion for OD and FER. Studies were excluded if they focused only on unimodal approaches, were unrelated to VIR fusion, lacked sufficient technical or evaluation details, or were outside the scope of OD and FER. The selected studies were finally organized according to dataset type, fusion level, application domain, evaluation metric, reported performance, limitations, and future research challenges. The overall literature search and study-selection workflow is shown in Figure 2.

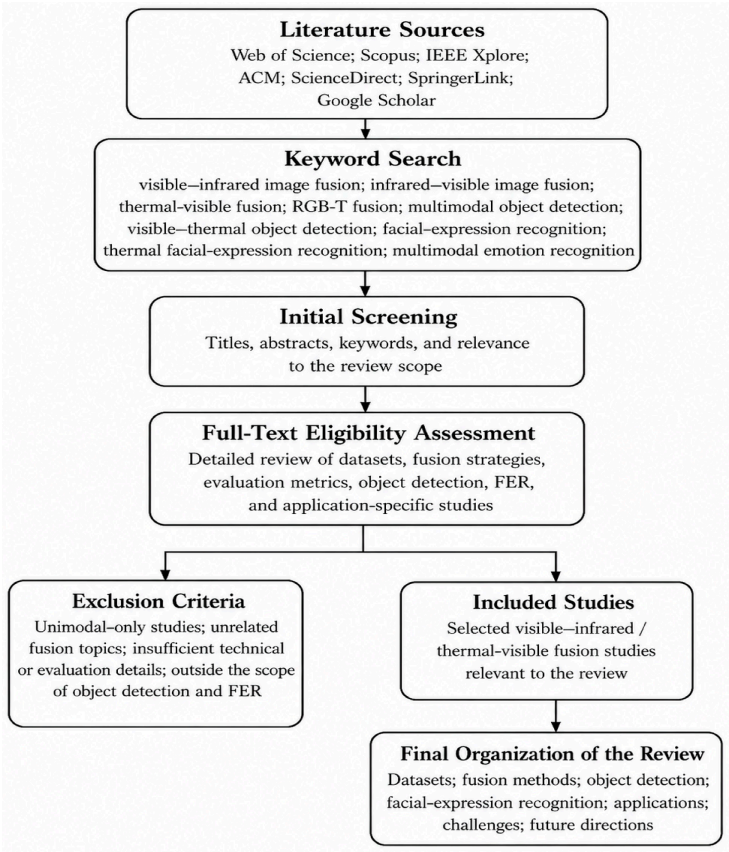

Figure 2. Literature search and study-selection workflow used in this review.

In addition, the publication years of the selected studies were analyzed to summarize the temporal development of VIR fusion research. The selected studies were grouped into four publication periods: before 2015, 2015–2018, 2019–2021, and 2022–2026. As shown in Figure 3, the number of selected studies has increased in recent years, indicating growing

research interest in VIR fusion and its applications in OD, FER, and related computer vision tasks.

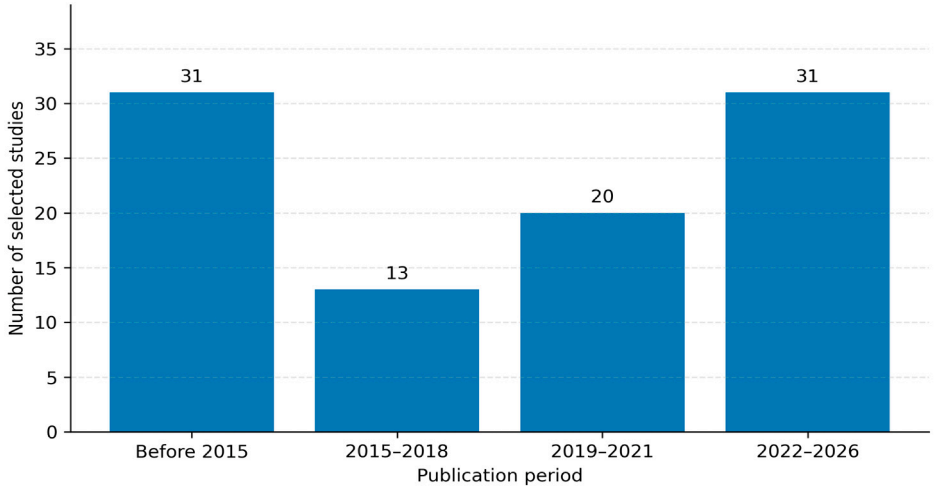

Figure 3. Publication trend of selected studies included in this review. The selected studies were grouped by publication period based on the publication years reported in the reference list; non-year dataset resources were excluded from the count.

The study contributes the following to the extant literature:

- It provides a task-oriented review of VIR fusion for two representative downstream computer vision tasks: OD as a relatively mature scene-level application and FER as an emerging human-centered application.
- It organizes VIR and related RGB-T datasets according to their primary task types, including image fusion, OD, tracking, registration, surveillance, and FER, thereby avoiding conflation of detection datasets with related RGB-T benchmarks.
- It compares early- and late-fusion strategies, including sensor-/pixel-level, feature-level, score-level, and decision-level fusion, with emphasis on their relevance to downstream OD and FER performance.
- It distinguishes task-specific evaluation metrics used in image fusion, OD, tracking, registration, and FER, and explains why results across these tasks should not be directly ranked using a single metric.
- It provides a cross-study synthesis of fusion-level strengths, limitations, and deployment challenges, including alignment sensitivity, dataset dependence, computational complexity, real-time constraints, and limited cross-dataset validation.
- It identifies OD- and FER-specific future directions, including alignment-robust detection, lightweight real-time fusion, standardized VIR benchmarks, larger paired VIR-FER datasets, demographic diversity, and ethical deployment of FER systems.

Compared with existing image-fusion surveys, the present review is distinguished by its task-oriented focus on VIR fusion for OD and FER. Previous surveys have mainly emphasized general image-fusion theory, medical image fusion, remote sensing fusion, VIR fusion algorithms, or neural network-based fusion methods. In contrast, this review connects VIR datasets, fusion strategies, task-specific evaluation metrics, OD-oriented studies, FER-oriented studies, cross-study synthesis, applications, and future challenges within one computer vision-oriented framework. This organization makes it possible to compare a relatively mature scene-level task, OD, with an emerging human-centered task, FER, and to identify both shared and task-specific limitations.

2. Multimodal Datasets

Robust datasets and reliable evaluation metrics are crucial to developing and assessing relevant algorithms in the field of VIR image fusion for OD and FER. These components are essential for fusion-technique training, testing, and benchmarking, as they provide a basis for evaluating the conventional methods and advancing the field. In this section, we first review VIR datasets and related RGB-T benchmarks used in fusion, OD, tracking, registration, and surveillance studies, and then analyze datasets used for FER.

2.1. Visible–Infrared Datasets and Related RGB-T Benchmarks

Here, we review VIR datasets and related RGB-T benchmarks used in image fusion, OD, tracking, registration, and surveillance studies. These datasets are also presented in Table 1. A previous study [27] introduced the TNO multiband image data collection, which is primarily an image-fusion and surveillance-scene dataset rather than a standard OD benchmark. The dataset includes multiband imagery acquired across the visible band (approximately 390–700 nm), near-infrared (NIR; approximately 700–1000 nm), and long-wave infrared (LWIR; approximately 8–12 μm) spectral ranges. These images were collected from military- and surveillance-related scenarios and are useful for evaluating image-fusion methods under low-light and poor-visibility conditions. The Kayak image-fusion sequence includes registered optical, NIR, and LWIR image sequences showing three kayaks, whereas the TRICLOBS dynamic multiband image collection captures motion sequences in the visible (400–700 nm), NIR (700–1000 nm), and LWIR (8–14 μm) bands for dynamic monitoring of urban environments. Therefore, in this review, TNO is treated as a multiband fusion resource for image-fusion and visibility-enhancement evaluation, not as a dedicated detection benchmark.

Xu et al. [28] released the recently aligned VIR image dataset, known as RoadScene. This dataset first selects image pairs from videos with high repetition rates, thereby suppressing thermal noise in the original IR images. The photo pairings are aligned using homologous and bicubic interpolation pairs, following careful feature-point selection; thereafter, the exact registration area is trimmed. The interpolation process, known as cubic interpolation, projects continuous pixel values onto a discrete pixel grid by computing a weighted average of neighboring pixels at each pixel position. Using a 16-pixel window, the bicubic interpolation algorithm subsequently calculates the value of the target pixel by averaging the 16 surrounding pixels, each weighted by the pixel's position within the window. This computation method is referred to as bicubic interpolation because it uses a cubic interpolation function. Additionally, bicubic interpolation was preferred, as it reduces distortion and artifacts across images and improves smoothness and image quality, especially when aligned across image pairs. This is essential for accurate feature matching and VIR image fusion. The RoadScene dataset comprises 221 image pairs, featuring intricate scenarios, including congested roads, pedestrians, and moving vehicles. Moreover, it addresses the lack of comprehensive information on IR images, the low spatial resolution, and the small number of image pairs in the reference dataset.

The Microsoft Common Objects in Context (MS-COCO) dataset [29] is a large-scale visible-image OD and segmentation dataset. It is not a paired VIR dataset and is therefore not included in Table 1 as a VIR benchmark. However, some visible–infrared object-detection studies use MS-COCO for visible-image pretraining, baseline model development, or domain-adaptation experiments because of its large number of annotated visible images.

Table 1. Summary of VIR datasets and related RGB-T benchmarks used in fusion, OD, and tracking studies.

Datasets	Primary Task/Type	Description
TNO [27]	Multiband image fusion/ surveillance scenes	Includes visible-band, NIR, and LWIR images from military and surveillance-related scenes; mainly used for multiband image-fusion and visibility-enhancement evaluation rather than as a standard OD benchmark.
RoadScene [28]	Image fusion/road scenes	Includes 221 image pairings of scenarios that are repeated extensively.
MSRS [30]	Image fusion/segmentation-related	Includes 1444 pairs of perfectly aligned visible and IR images.
LLVIP [31]	Low-light pedestrian detection	Includes 15,488 strictly time- and space-aligned image pairings, the majority of which were captured in dimly lit environments.
M3FD [32]	VIR fusion and OD	Contains 4200 aligned VIR image pairs with object annotations for six categories, supporting image fusion and OD evaluation.
DroneVehicle [33]	RGB–IR aerial vehicle detection	Contains 28,439 paired RGB–IR aerial images with vehicle annotations, supporting RGB–infrared vehicle detection under diverse aerial-view conditions.
RGBT210 [34]	RGB-T tracking benchmark	Includes 210 video sets that are pixel-by-pixel aligned.
RGBT234 [35]	RGB-T tracking benchmark	Includes 234 video sets, increasing scene diversity.
OTCBVS [36]	Surveillance/tracking/ fusion benchmark	Includes 14 sub-datasets, including one visible-light dataset, seven IR datasets, and six VIR datasets.
LITIV [37]	Registration/tracking/surveillance	Comprises several sub-data sets and a wide range of distinct sensor data.
GTOT [38]	RGB-T tracking benchmark	Includes fifty films depicting various scenes, including LABS, swimming pools, and roadways.
INO [39]	Video analytics/object scenes	Includes 2000 images in 10 different categories, each measuring 256 × 256 pixels.
FLIR [40]	Thermal/RGB OD	Consists of 14,452 annotated thermal images, 10,228 of which were taken from short films and 4224 from a 144 s continuous video.
Marine [41]	Maritime OD	Includes paired visible and IR maritime samples with annotated vessels: Scenario 1 contains 7250 training pairs and 1750 test pairs acquired during daytime, totaling 9000 paired samples; Scenario 2 contains 2250 training pairs and 1000 test pairs acquired during nighttime, totaling 3250 paired samples.
KAIST [42]	Multispectral pedestrian detection	Includes 95,328 color-thermal image pairs with detailed annotations, including temporal correspondences and occlusion tags.

The MSRS dataset is a novel multispectral collection of 1444 matched image pairs [30], providing a resource for multisensory image processing research. They gathered 715 daytime and 729 nighttime image pairs from the MFNet dataset, removed 125 unaligned pairs, and applied dark channel image enhancement to improve contrast and signal-to-noise ratio. The LLVIP dataset [31] contains 15,488 image pairs and 30,976 images captured in dark conditions, with precise time and space alignment. The dataset includes pedestrian images taken in low light, where IR images complement visible images. The images are high quality, with 1920 × 1080 visible light and 1280 × 720 IR resolution. LLVIP's advantages make it valuable for pedestrian detection in low-light conditions and will advance computer vision through image fusion and conversion applications.

The M3FD dataset is a recent VIR benchmark designed for image fusion and OD [32]. It contains 4200 aligned visible and infrared image pairs collected under different driving conditions, including daytime, overcast, nighttime, and challenging scenes. The dataset includes annotations for six object categories, including person, car, bus, motorcycle, truck, and traffic light, making it suitable for evaluating both fusion quality and downstream detection performance. The DroneVehicle dataset is a large-scale drone-based RGB–IR vehicle-detection benchmark [33]. It contains 28,439 RGB–IR image pairs collected from urban roads, residential areas, parking lots, and other scenarios from daytime to nighttime. The dataset provides vehicle annotations and is particularly useful for evaluating RGB–infrared fusion and detection under aerial-view and low-light conditions.

In [34], researchers captured the RGBT210 dataset using a CCD camera (SONY EXView HAD CC) and an imager (DLS-H37 DM-A) [35]. The cameras were aligned at the pixel level, sharing image characteristics with a parallel optical axis across the collimator. The RGBT234 dataset [35] adds video sequences captured under hot-weather conditions and improves scenery diversity compared to RGBT210. The dataset comprises 234 video sets with approximately 233,800 frames. The OTCBVS dataset [36] is a public benchmark for algorithm evaluation, containing images of faces, gestures, weapons, ships, and people. It

includes 14 sub-datasets, including one visible-light dataset, seven IR datasets, and six VIR datasets. The LITIV dataset [37] is a multimodal resource for computer vision research, containing diverse images and annotations for OD, segmentation, and tracking. It includes sensor data like images, videos, IR images, and depth images across multiple sub-datasets. The GTOT dataset [38] contains 15,800 frames across 50 contexts, featuring various objects such as swans and cars. The INO dataset [39] consists of 2000 256 × 256 pixel images organized into 10 groups, each representing a different object or scene.

The FLIR Thermal Starter Dataset [40] provides thermal and RGB images for OD training. It includes 4224 images from a 144 s video and 10,228 images from short videos, totaling 14,452 annotated thermal images. The data were collected using car-mounted thermal and RGB cameras in Santa Barbara, California, on public roads during the day and at night. The researchers in [41] collected a maritime dataset in the Finnish archipelago using visible spectrum and IR cameras. The recordings were synchronized to create panoramic views. The thermal cameras had VGA resolution with a 35° horizontal field of view, while the visible cameras had full HD resolution. Manual registration ensured aligned images with a one-frame-per-second sampling rate. The dataset contains paired visible and IR maritime samples with annotated vessels, featuring two scenarios: Scenario 1 includes 7250 training pairs and 1750 test pairs acquired during daytime, totaling 9000 paired samples; Scenario 2 includes 2250 training pairs and 1000 test pairs acquired during nighttime near the harbor, totaling 3250 paired samples. The KAIST multispectral pedestrian dataset was collected using a car-mounted system [42] comprising a color camera (PointGrey™ Flea3, Point Grey Research Inc., Richmond, BC, Canada) and thermal camera (FLIR-A35, FLIR Systems, Inc., Wilsonville, OR, USA) with a beam splitter for optical alignment. The dataset includes 95,328 color–thermal pairs with annotations, collected in various traffic scenarios under different lighting conditions. A three-axis camera jig and special calibration board ensured precise alignment between color and thermal images. Because several RGB-T resources reviewed in this study were originally designed for tracking or registration rather than OD, Table 1 identifies the primary task/type of each dataset to avoid conflating detection datasets with related RGB-T benchmarks.

2.2. Multimodal Datasets for FER

The available multimodal datasets used for FER are listed in Table 2. Among the numerous datasets, the Visible and Infrared Image Fusion (VIRI) dataset [15] is an important resource that provides naturally occurring images from uncontrolled environments. The dataset, generated at the University of Toledo, includes five expressions: happy, sad, surprise, anger, and neutral. VIRI was created with 110 subjects, including 70 male and 40 female participants, from different demographic backgrounds, including Asian, African American, and American participants. Participants were mostly on-campus students aged 17–35. Other specialized datasets exist for applications like surveillance, automotive navigation, and medical imaging, supporting algorithm development and technique assessment. The NIST dataset [43] contains 1919 IR images of 600 people showing three expressions: surprise, frown, and smile. The NVIE [44] dataset includes visible and IR images in spontaneous and posed categories, captured under three lighting settings for 215 people (157 male, 58 female) aged 17–31. Using online video snippets, subjects displayed six expressions: happy, sad, surprise, fear, anger, and disgust, along with emotions elicited by the videos.

Table 2. Summary of multimodal VIR datasets used for FER.

Datasets	Subjects	Spectrum Range	Description
VIRI [15]	110	8–14 μm	A spontaneous dataset from 110 subjects comprising paired visible and IR images, consisting of five expressions (happy, sad, surprise, anger, and neutral).
NIST [43]	600	8–12 μm 3–5 μm	A posed dataset from 600 subjects consisting of three expressions (smile, frowning, and surprise).
NVIE [44]	215	8–14 μm	A posed and spontaneous dataset from 215 subjects consisting of six expressions (happy, sad, surprise, fear, anger, disgust, and neutral).
KTFE [45]	26	8–14 μm	A spontaneous dataset from 26 subjects consisting of seven expressions (happy, sad, surprise, fear, anger, disgust, and neutral).
KTFEv2 [46]	30	8–14 μm	A posed and spontaneous dataset from 30 subjects consisting of seven expressions (anger, disgust, fear, happy, sad, surprise, and neutral).
IRIS [47]	30	7–14 μm	A posed dataset from 30 subjects consisting of three expressions (surprise, laughter, and anger).

The KTFE [45] is a thermal FER dataset containing seven expressions (happy, sad, surprise, fear, anger, disgust, and neutral) and spontaneous expressions. The corpus includes 26 people aged 11–31 from Vietnam, Japan, and Thailand, featuring visual and thermal videos. As in the NVIE dataset, subjects were presented with emotive excerpts as stimuli. The KTFE dataset is divided into four categories based on emotion intensity. However, the thermal files require proprietary software to view, making it difficult to use.

KTFEv2, the second-generation dataset [46], contains visual and thermal data from 30 subjects aged 11–32 from Thailand, China, Japan, and Vietnam, mostly JAIST students. Researchers used multifunctional software to select expressions, with three people selecting optimal frames for each emotion and two people retrieving images from the dataset. KTFEv2 categorizes expressions into three intensity levels (low, medium, and high) to meet real-world needs.

The IRIS dataset [47] contains visible and thermal face images in various lighting conditions at 320 × 240 resolution. It features 30 subjects (28 male, two female) expressing surprise, laughter, and anger, with 176–250 images per person and 11 images/rotations each. The dataset uses five illumination settings to produce 4228 pairs of visual and thermal images.

3. Data Fusion Methods

In VIR image fusion, visible images mainly provide texture, edge, color, and appearance information, whereas IR images provide thermal contrast and target-saliency cues that remain useful under low illumination, adverse weather, smoke, haze, or partial occlusion. Therefore, VIR fusion differs from generic multisensory fusion because it must address modality-specific challenges, including spatial misalignment, different sensor resolutions, heterogeneous spectral responses, thermal noise, loss of fine texture in IR images, illumination sensitivity in visible images, and possible temporal synchronization errors. These characteristics make the choice of fusion level particularly important for downstream computer vision tasks such as OD and FER.

In this review, sensor-/pixel-level fusion, feature-level fusion, score-level fusion, and decision-level fusion are treated as distinct concepts. Sensor-/pixel-level fusion refers to early fusion performed before substantial feature extraction, where raw, minimally

processed, or registered image-level visible and IR information is combined. Feature-level fusion combines modality-specific representations extracted from visible and IR branches, usually within a machine learning or deep learning architecture. Score-level fusion combines confidence scores or class probabilities generated by modality-specific models, whereas decision-level fusion combines final predictions using majority voting, weighted voting, rule-based selection, or other decision-integration strategies. This distinction is important because the computational cost, alignment requirement, interpretability, and task dependence differ across fusion levels. In this review, pixel-level fusion refers to combining registered visible and IR images before feature extraction to produce a fused image. Feature-level fusion refers to combining intermediate modality-specific feature maps or embeddings inside a learning model. Score-level fusion refers to combining model outputs such as confidence scores, class probabilities, logits, or softmax-level outputs before the final decision. Decision-level fusion refers to combining final predictions, such as detected object labels, bounding boxes, expression classes, or modality-specific decisions, using voting, weighting, or rule-based integration. Thus, the distinction between early and late fusion is based on the exact dataflow stage at which pixels, features, scores/logits, or final decisions are combined. The process of combining data from several sources to create more accurate, trustworthy, and thorough information is known as data fusion. This process improves the data analysis and decision-making procedures. Data fusion methods can be broadly categorized into early fusion and late fusion, each with its own subcategories [48], as depicted in Figure 4.

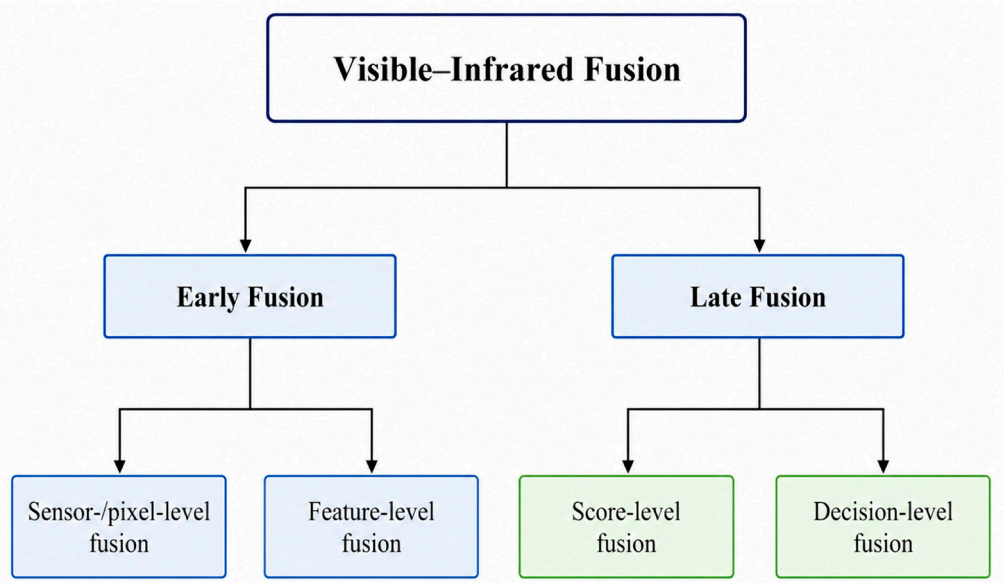

Figure 4. Taxonomy of VIR fusion strategies, distinguishing early fusion approaches, including sensor-/pixel-level and feature-level fusion, from late fusion approaches, including score-level and decision-level fusion.

VIR studies can also be distinguished according to whether they follow a fusion-then-detect strategy or a detection-oriented fusion strategy. In fusion-then-detect approaches, VIR images are first fused into a single image, and a detector or classifier is then applied to the fused output. This strategy is relatively interpretable and allows the fused image to be assessed using image-quality metrics, but the fusion objective may not always be optimized for the downstream task. In contrast, detection-oriented fusion integrates visible and IR information inside the detection or recognition network and optimizes the fusion process together with task-specific losses, such as classification or bounding-box regression losses. This strategy is often more task-adaptive and can improve detection robustness,

but it usually requires paired multimodal training data, accurate alignment, and higher computational resources. Therefore, the choice between fusion-then-detect and detection-oriented fusion depends on whether the priority is fused-image quality, downstream accuracy, interpretability, or real-time deployment.

3.1. Early Fusion

3.1.1. Sensor-/Pixel-Level Fusion

Sensor-/pixel-level fusion combines registered VIR images before feature extraction, as summarized in Figure 5a. In this strategy, the visible and IR inputs are first aligned, and then a fusion module generates a fused image that is subsequently used by a detector or classifier [48]. This approach is commonly associated with fusion-then-detect pipelines because the fusion process is completed before the downstream task model is applied. Its main advantage is interpretability, as the fused image can be visually inspected and evaluated using image-quality metrics. However, it is highly sensitive to spatial misalignment, sensor-resolution differences, and fusion artifacts, which can reduce downstream detection or recognition performance.

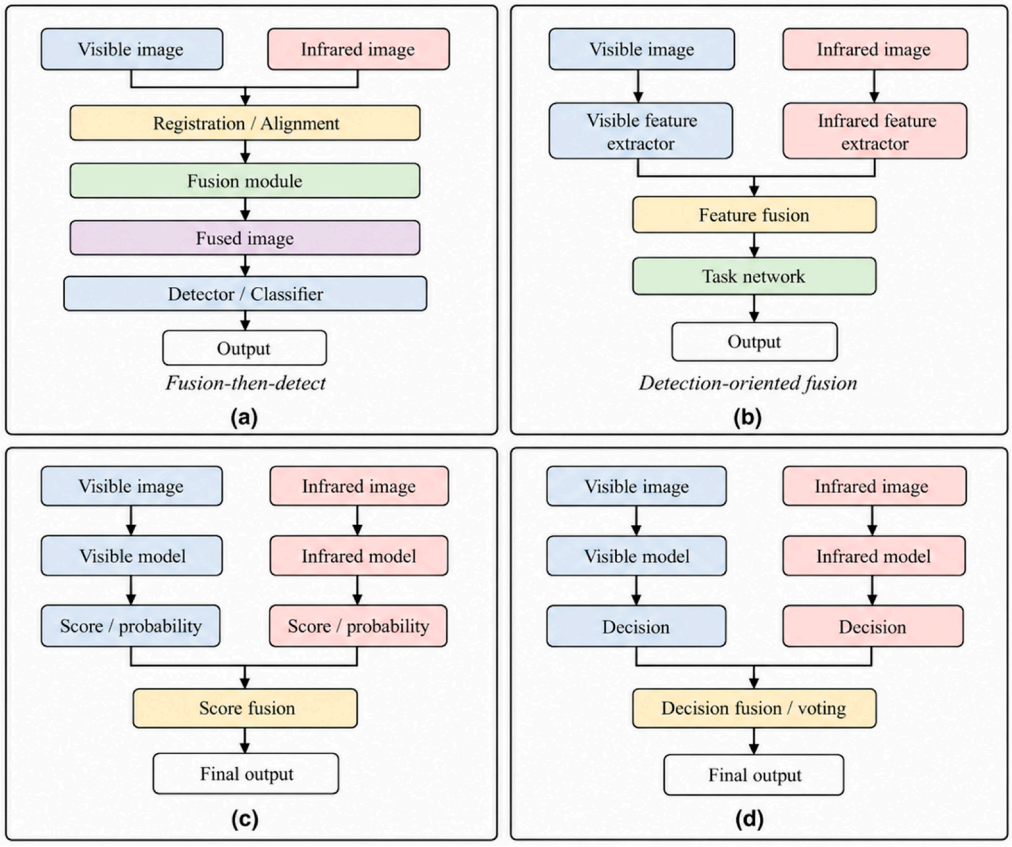

Figure 5. Comparative workflows of VIR fusion strategies for downstream computer vision tasks. (**a**) Sensor/pixel-level fusion combines registered VIR images before feature extraction and typically follows a fusion-then-detect strategy. (**b**) Feature-level fusion combines modality-specific representations inside the model and is commonly used in detection-oriented fusion. (**c**) Score-level fusion combines confidence scores from modality-specific models. (**d**) Decision-level fusion combines final predictions using voting or rule-based integration.

3.1.2. Feature-Level Fusion

Feature-level fusion combines modality-specific representations extracted from visible and IR branches, as summarized in Figure 5b. In this strategy, each modality is processed by a separate feature extractor, and the resulting representations are merged using concate-

nation, attention, weighted summation, transformer-based interaction, or other learnable fusion mechanisms [48]. Feature-level fusion is commonly used in detection-oriented fusion because the fusion process can be optimized together with task-specific objectives. Compared with sensor-/pixel-level fusion, it can better preserve modality-specific information, but it usually requires paired multimodal data, careful network design, and greater computational resources. Recent approaches focused on fusion algorithms based on neural network models can be found in [25]. Unlike these broader fusion-method surveys, the present review uses fusion-level categories to interpret how VIR fusion affects downstream OD and FER performance, rather than only summarizing fusion algorithms or image-quality improvement.

3.2. Late Fusion

3.2.1. Score-Level Fusion

Score-level fusion combines confidence scores or class probabilities generated by modality-specific models, as summarized in Figure 5c. In this strategy, visible and IR inputs are processed independently, and their output scores are combined using weighted averaging, max pooling, probabilistic fusion, or learned score-integration rules [48]. Score-level fusion is useful when modality-specific models are already available or when separate model outputs need to be integrated without modifying internal feature representations. However, it may be less adaptive than feature-level fusion because the interaction between visible and IR information occurs only at the output-score stage.

3.2.2. Decision-Level Fusion

Decision-level fusion combines final predictions produced independently from visible and IR models, as summarized in Figure 5d. The final decision can be obtained using majority voting, weighted voting, rule-based selection, or confidence-based decision integration [48]. This strategy is simple and modular because each modality can be trained and evaluated separately. However, because fusion occurs only after individual decisions have already been made, decision-level fusion may fail to exploit fine-grained complementary information between visible texture cues and IR thermal cues.

Each fusion method has advantages suited to different applications, depending on the nature of the data, desired outcome, and task requirements. Sensor-/pixel-level and feature-level fusion are used when image-level data or intermediate features provide complementary insights. Score- and decision-level fusion are employed when combining model results could yield more accurate final outcomes.

4. Application-Specific Studies of Visible–Infrared Fusion

In this section, we first discuss performance measures used in VIR image-fusion studies, then examine application-specific studies using multimodal datasets for OD and FER.

4.1. Evaluation Metrics

To objectively evaluate VIR fusion and its downstream computer vision performance, two groups of metrics are commonly used: image-fusion quality metrics [49,50] and task-performance metrics. Image-fusion quality metrics assess the information content, contrast, structure, edge preservation, and visual fidelity of fused images, whereas task-performance metrics evaluate OD, tracking, or FER performance. The following metrics are commonly reported in the reviewed studies:

- **Entropy (EN):** Determines how much information an image has. A higher entropy following fusion signifies better fusion performance (Equation (1)).

- **Average Precision (AP):** Measures the precision–recall trade-off in OD and information retrieval by averaging the precision values at various recall levels. AP shows how well the model balances recall and precision (Equation (2)).
- **Mean Average Precision (mAP):** Extends AP to evaluate models across multiple classes by averaging the AP scores for each class, providing a comprehensive performance measure (Equation (3)).
- **Accuracy:** Evaluates the performance of a recognition model (Equation (4)).
- Recall: Evaluates how many real positive cases the model accurately recognized, which is important for datasets that are not balanced (Equation (5)).
- **Success Rate (SR):** Calculates the proportion of successful outcomes from all trials or attempts (Equation (6)).
- **Mutual Information (MI):** Determines how much information is moved from the original images to the combined image (Equation (7)).
- **Structural Similarity Index (SSIM):** Evaluates structural similarity between images by considering luminance, contrast, and structural information. It is commonly used to assess whether the fused image preserves structural details from the source images.
- **Visual Information Fidelity (VIF):** Estimates the amount of visually meaningful information preserved in the fused image.
- **Edge-preservation metric (Qabf):** Measures how effectively edge strength and orientation information from the source visible and IR images are transferred to the fused image.
- **Standard Deviation (SD):** Reflects the contrast and gray-level dispersion of the fused image. A higher SD generally indicates stronger image contrast.
- **Spatial Frequency (SF):** Measures image activity and spatial detail. A higher SF indicates richer texture and sharper spatial variation.
- **Correlation Coefficient (CC):** Measures the statistical similarity or information consistency between the fused image and the source images:

$$\mathbf{EN} = -\sum_{\mathbf{n=0}}^{\mathbf{N-1}} \mathbf{p_n log_2 p_n} \tag{1}$$

where **N** stands for the fused image's gray level, and $\mathbf{p_n}$ denotes the normalized histogram of the corresponding gray level:

$$\mathbf{AP} = \sum_{\mathbf{n=1}}^{\mathbf{N}} (\mathbf{R_n} - \mathbf{R_{n-1}})\mathbf{P_n} \tag{2}$$

where P_n is the precision at the n-th threshold, R_n is the recall at the n-th threshold, and N is the total number of thresholds where precision and recall values change:

$$\mathbf{mAP} = \frac{\mathbf{1}}{\mathbf{C}} \sum_{\mathbf{i=1}}^{\mathbf{C}} \mathbf{AP_i} \tag{3}$$

where AP_i is the AP for the i-th class, and C denotes the number of classes:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

where FN indicates how many positive cases were mistakenly labeled as negative, and TP indicates how many positive cases were accurately classified as positive:

$$\text{SR} = \frac{\text{Number of Successful Outcomes}}{\text{Total Number of Attempts}} \tag{6}$$

$$\mathbf{MI} = \sum_{x=0}^{L-1} \sum_{y=0}^{L-1} \mathbf{P_{AB}}(x, y) \log_2 \frac{\mathbf{P_{AB}}(x, y)}{\mathbf{P_A}(x)\mathbf{P_B}(y)} \quad (7)$$

Selecting appropriate evaluation metrics is critical, as applications prioritize different aspects of image quality. Surveillance and biometric applications often prioritize thermal information for identifying subjects under low-light or poor-visibility conditions [37,43], whereas medical imaging focuses on clarity of anatomical structures [51]. Metrics should align with the intended use case to ensure practical utility. However, objective metrics have limitations, as they may not fully capture human observers' subjective experience [52]. Therefore, developing comprehensive datasets and refining evaluation metrics are essential for advancing visible and IR image fusion [53]. Creating diverse datasets and developing perceptually relevant metrics will be key to advancing image fusion technology [52].

Because the reviewed studies address different tasks, datasets, splits, backbones, and evaluation protocols, their reported metrics are not directly interchangeable. For clarity, the metrics used in the reviewed studies can be divided into two broad groups: image-fusion quality metrics and downstream task-performance metrics. Image-fusion quality metrics evaluate the information content, contrast, structure, edge preservation, and visual fidelity of the fused image. These include entropy (EN), mutual information (MI), structural similarity index (SSIM), visual information fidelity (VIF), edge-preservation metric (Qabf), standard deviation (SD), spatial frequency (SF), and correlation coefficient (CC). These metrics are useful for assessing fused-image quality, but they do not directly measure OD, tracking, or FER performance.

In contrast, downstream task-performance metrics evaluate how well the fused or multimodal representation supports a specific computer vision task. OD studies usually report average precision (AP), mean average precision (mAP), precision, and recall. Tracking studies may report success rate (SR), whereas FER studies mainly report classification metrics such as accuracy, precision, recall, and F1-score. Therefore, the results reported in Tables 3 and 4 should be interpreted only within their respective task categories and experimental settings. They are provided to summarize reported outcomes from individual studies, not to rank methods across different datasets, tasks, protocols, or metric types. For example, EN values from image-fusion studies should not be directly compared with AP, mAP, F1-score, success rate, or relative-improvement values reported by detection, tracking, or recognition studies.

4.2. Studies on Object Detection with Visible and IR Images

In OD, multimodal datasets are essential for developing and evaluating fusion methods that combine complementary visible and IR information for target localization and recognition. In this subsection, the reviewed OD-related studies are organized according to the stage at which visible and IR information is combined, rather than being treated as a simple chronological catalogue of individual papers. Some methods follow a fusion-then-detect pipeline, where a fused image is first generated and then used for OD. Other methods perform detection-oriented fusion by combining visible and IR features inside the detection network. A smaller group of studies combines modality-specific scores or final decisions at later stages. Therefore, the methods summarized in Table 3 are interpreted according to task type, fusion stage, dataflow, reported metric, and practical limitation, rather than as directly comparable or mutually exclusive approaches. Although the main focus of this subsection is OD, a small number of tracking and registration studies are also included because they use VIR or RGB-T data and report fusion strategies that are technically relevant to detection-oriented multimodal perception. These studies are not treated as standard object-detection benchmarks; instead, they are discussed as related RGB-T perception studies that provide useful evidence about image alignment, modality

interaction, and fusion-level design. To avoid conflating different tasks, Table 3 identifies the primary task type and the reported evaluation metric for each study. To avoid conflating different fusion objectives, the reviewed OD studies are further distinguished according to what is fused and where fusion occurs in the architecture. In fused-image generation methods, the visible and IR inputs are combined at the image or pixel level to produce a discrete fused image, which can be evaluated using image-quality metrics such as EN, MI, SSIM, or related fusion measures. In detection-oriented feature-fusion methods, the visible and IR streams are usually processed by separate or partially shared network branches, and their intermediate feature maps are combined inside the detector using concatenation, attention, selection, transformer interaction, or other learnable fusion modules. These methods are primarily evaluated using detection metrics such as AP or mAP because their main objective is to improve classification confidence and bounding-box localization. Some task-driven methods, such as TarDAL, lie between these two categories because they generate a fused image while also optimizing detection-related objectives. Therefore, image-quality metrics and detection metrics are reported as task-specific outcomes and are not interpreted as directly interchangeable. Within fusion-then-detect and joint fusion–detection methods, semantic and attention-based fusion strategies have been used to connect fused-image generation with downstream detection performance. For example, the multilevel structured search attention fusion network in [54] uses semantic information, neural architecture search, and a multilevel adaptive attention module to preserve useful modality-specific cues while suppressing redundant fusion information. By linking the fusion network with detection-related training objectives, this type of method attempts to make the fused representation more useful for object classification and detection, rather than optimizing only visual image quality. However, its effectiveness still depends on accurate semantic guidance, reliable modality alignment, and the consistency between the fusion objective and the downstream detection task.

Task-driven fusion- and detection-oriented learning methods further show how VIR-related models can be optimized for downstream OD rather than only for image enhancement. TarDAL [55] generates a fused image using TarDAL while jointly optimizing detection-related objectives to improve downstream OD performance. It should be noted that task-driven fusion methods, such as TarDAL, differ from architectures that only consume visible and IR inputs inside a detector. TarDAL produces an explicit fused image and simultaneously uses detection-related optimization to make the fused output useful for downstream OD. In contrast, detector-internal VIR architectures such as RSDet and M^2FNet mainly fuse intermediate visible and IR feature representations inside the detection network and are evaluated primarily through detection performance rather than standalone fused-image quality. Its fusion network uses a generator and two discriminators to preserve visible textural details and IR structural information, while the detection objective encourages the fused output to remain useful for bounding-box prediction. This makes TarDAL different from purely image-quality-oriented fusion methods because the generated fused image is evaluated not only visually but also through its contribution to OD performance. Other studies emphasize detection robustness without necessarily producing a standalone fused image. Contrastive-learning-based YOLOv5 [56] improves object-background discrimination by extracting object and background regions and optimizing contrastive loss together with the object-detection loss. Similarly, the domain-adaptive YOLOv5 framework in [57] addresses cross-domain and small-object detection by combining a domain classifier, Wasserstein-distance-based loss, knowledge distillation, and domain-confusion learning. These methods are relevant to VIR-oriented OD because they demonstrate how task-specific learning objectives can improve detection robustness under modality variation, domain

shift, and challenging imaging conditions; however, they should be distinguished from methods that explicitly generate fused VIR images.

Pixel-level and feature-level fusion studies show different ways of integrating visible and IR information for OD. The adaptive pixel-weighting strategy in [58] learns pixel-by-pixel fusion weights to generate fused images while also using joint optimization with a detection model. This places the method between image-level fusion and task-driven fusion, because it evaluates both fused-image quality and downstream detection accuracy. However, its performance can depend on image resolution, accurate registration, and the consistency between pixel-level fusion objectives and detection objectives. In contrast, VIFF [59], RSDet [60], and M^2FNet [61] are more representative of detection-oriented feature fusion, where visible and IR information is combined inside the detection network. VIFF [59] uses two modality-specific processing units: the IR branch enhances contrast and semantic content, whereas the visible branch preserves gradient and texture information; the extracted features are then integrated using attention mechanisms. RSDet [60] performs detector-internal feature fusion using coarse redundant-spectrum removal and dynamic feature selection, which is primarily optimized for OD performance. M^2FNet [61] performs detector-internal visible–thermal feature fusion using union-modal attention and cross-modal attention; primarily optimized using OD performance. These methods indicate that recent VIR-OD research is shifting from simply generating visually enhanced fused images toward learning task-adaptive cross-modal representations that directly improve detection robustness.

Recent detection-oriented VIR and RGB-X studies have further emphasized task-adaptive fusion rather than only image-quality-oriented fusion. Deevi et al. [62] proposed an RGB-X object-detection framework based on scene-specific fusion modules, showing that multimodal fusion can be adapted according to scene characteristics and auxiliary sensor modality. This type of modular fusion is useful for RGB–thermal and other RGB-X detection settings because it enables the network to selectively exploit complementary modality-specific information according to scene characteristics. Hou et al. [63] presented an infrared–visible dual-modal feature-fusion OD algorithm that integrates information from both modalities to improve target representation under complex imaging conditions. These recent studies indicate that VIR-OD is moving toward detection-oriented, feature-level, and adaptive fusion architectures, where the fusion process is optimized to improve downstream detection robustness rather than only to enhance visual quality. More recent RGB–IR and infrared–visible OD studies have continued this shift toward task-oriented fusion. For example, FQDNet uses a fusion-enhanced quad-head RGB–IR object-detection framework to improve multimodal feature integration and multi-scale detection [64], whereas Tan et al. proposed an infrared–visible multimodal detection method based on a cross-modal information bottleneck and minimum-redundancy transformation to reduce modality redundancy and improve feature alignment [65].

Other detection-oriented studies focus on improving robustness under low signal-to-noise conditions or combining heterogeneous fusion stages within a single detection framework. DBD-YOLOv8 [66] is mainly an IR OD method rather than a full VIR fusion approach. It addresses YOLOv8 limitations in low signal-to-noise IR images by using deformable convolution to adapt receptive fields, bi-level routing attention to improve semantic representation, and dynamic heads to refine multiscale feature maps. Therefore, its relevance to this review lies in showing how detection-specific architectural changes can improve IR-based target detection under degraded imaging conditions. MFDetection [67], in contrast, is more directly related to VIR fusion because it integrates pixel-level and feature-level fusion across heterogeneous visible and IR inputs within a unified detection framework. By combining multilayer image fusion with OD and sharing feature

extraction across stages, the model aims to improve detection accuracy while reducing redundant computation. This type of architecture illustrates a multi-stage fusion design, where image-level and feature-level information are both used to support downstream detection performance.

Registration, tracking, and late-fusion studies provide complementary evidence about the practical requirements of VIR-based OD. Accurate visible–IR alignment is a prerequisite for reliable pixel-level and feature-level fusion. The registration method in [68] addresses this issue by using OD to extract constrained point features, applying the left-value rule for strict point matching, and estimating the affine transformation between visible and IR images. Although this study is not a standard OD benchmark, it is relevant because misregistration can directly degrade fused-image quality and downstream detection performance. Related RGB-T tracking and detection studies further illustrate why some papers involve more than one fusion stage. The RGB-T tracking framework in [3] combines pixel-level, attention-based feature-level, and decision-level fusion to exploit complementary visible and thermal information across the tracking pipeline. This multi-stage design explains why such studies may appear in more than one fusion category in the synthesis table. In contrast, the YOLOX-based method in [69] represents late decision-level fusion, where visible and IR detections are combined at the output stage to improve low-light OD. The Faster R-CNN with feature pyramid network approach in [70] represents detection-oriented multispectral feature fusion, where RGB and thermal features are integrated to improve nighttime detection. Together, these studies show that alignment quality, fusion stage, and detector architecture jointly influence multimodal OD performance.

Deployment-oriented VIR studies show how fusion can support OD in practical systems such as ADAS, surveillance, and embedded smart platforms. In ADAS settings, fused visible–IR images have been used with DL-based detectors to improve object-recognition performance under low-light or poor-visibility conditions [71]. This type of approach highlights the importance of both image alignment and fusion quality, because misaligned visible and IR inputs can reduce the reliability of downstream detection. For embedded and low-power systems, fast multispectral fusion networks based on SSD-style CNN architectures have been explored to support real-time pattern recognition [72]. These studies indicate that practical VIR-OD deployment requires not only accuracy improvement but also attention to alignment, inference speed, hardware constraints, and robustness under real-world operating conditions. To better reflect recent progress and task diversity, Table 3 includes earlier fusion-based detection studies, recent detection-oriented VIR/RGB-X methods, and a limited number of related RGB-T tracking or registration studies that provide relevant insights into multimodal fusion, alignment, and perception.

Table 3. Summary of VIR fusion studies for OD and related RGB-T perception tasks.

Study	Task Type	Architecture	Datasets	Reported Metric/Task Outcome	Proposed Method	Limitations
Liu Yong et al. [54]	Image fusion + OD	MAAB	TNO RoadScene	EN: 7.310 EN: 7.365	Multilevel attention fusion network	Needs semantic guidance and accurate alignment; attention search may overfit on limited scenes.
Liu Jinyuan et al. [55]	Image fusion + OD	TarDAL network with one generator and dual discriminators	TNO RoadScene	EN: 2.766 EN: 3.378	Applies bi-level optimization with TarDAL to preserve visible details and IR structures.	Requires calibrated image pairs; adversarial training may weaken small thermal targets.
Tu et al. [56]	OD	YOLOv5	MS-COCO (visible-image pretraining/baseline; not paired VIR)	mAP: 58.7	Adds contrastive learning to YOLOv5 for better object-background distinction and detection.	Sensitive to object/background crop quality and visible–thermal domain mismatch.

Table 3. *Cont.*

Study	Task Type	Architecture	Datasets	Reported Metric/Task Outcome	Proposed Method	Limitations
Kim et al. [57]	IR OD/ domain adaptation	YOLOv5	MS-COCO (visible-image pretraining/baseline; not paired VIR)	mAP: 64.7	Enhances YOLOv5 with domain adaptation, auxiliary classifier, Wasserstein loss, and synthetic datasets.	Depends on synthetic/source-domain data; domain shift and thermal noise may reduce transferability.
Zhang et al. [58]	Image fusion + OD	YOLOv5	MSRS	$mAP_{0.5}$: 0.752	Pixel-based fusion network with adaptive weights, jointly optimized with detection model.	Sensitive to misregistration and resolution mismatch; pixel weighting may suppress weak thermal cues.
Yang et al. [59]	VIR-OD	VIFF	LLVIP	AP_{50}: 0.604	Uses dual processing units and attention mechanisms to fuse features	Attention fusion adds computation; downsampling may reduce fine thermal details.
Zhao et al. [60]	RGB-IR OD	RSDet	LLVIP FLIR	mAP: 61.3 mAP: 41.4	Uses RSDet with coarse removal and fine selection to fuse features.	Incorrect spectral removal may discard useful weak thermal or visible cues.
Deevi et al. [62]	RGB-X OD	RGB-X OD network with scene-specific fusion modules	RGB-X/RGB–thermal detection datasets	Detection metrics such as AP/mAP	Uses scene-specific fusion modules to adaptively combine RGB and auxiliary modality features for OD.	Requires diverse scene-specific training; performance depends on auxiliary-modality quality.
Hou et al. [63]	VIR-OD	VIR dual-modal feature-fusion detector	VIR-OD datasets	Detection metrics such as AP/mAP	Uses dual-modal feature fusion to improve object representation and detection under complex imaging conditions.	Sensitive to modality misalignment and cross-modal noise in complex scenes.
Meng et al. [64]	RGB-IR OD	FQDNet with fusion-enhanced quad-head detection	RGB-IR OD datasets	Detection metrics such as AP/mAP	Uses an optimized RGB-IR fusion strategy with a quad-head detection framework to improve multimodal feature integration and multi-scale detection.	Requires reliable RGB-IR alignment; multi-head design may increase computational cost.
Tan et al. [65]	VIR-OD	Cross-modal information bottleneck and minimum-redundancy transformation framework	VIR OD datasets	Detection metrics such as AP/mAP	Uses cross-modal information bottleneck learning and minimum-redundancy transformation to reduce modality redundancy and improve feature alignment for infrared–visible OD	Performance depends on effective modality alignment and redundancy suppression; additional modules may increase model complexity.
Jiang et al. [61]	Visible–thermal OD	M^2FNet with Transformer-based fusion	Visible and thermal IR datasets	mAP improvement: 10.71% over visible-only, 2.97% over thermal-only; up to 25.6% improvement under low-light conditions	Uses union-modal attention and cross-modal attention to fuse visible and thermal IR features for robust OD under different illumination conditions.	Transformer attention improves interaction but increases memory/computation for high-resolution inputs.
Shen et al. [66]	IR OD	Yolov8	OTCBVS	mAP: 81.6%	Enhances YOLOv8 with deformable convolutions, bi-level attention, and dynamic heads.	IR-only design; weak thermal contrast and small objects remain challenging.
Peng et al. [67]	Image fusion + OD	MF Detection	INO	mAP: 0.72	MF Detection fuses multiscale VIS and IR features for image fusion and OD.	Multilevel fusion is architecture-heavy and sensitive to misregistration/resolution gaps.
Li et al. [68]	VIR registration	CNN	LITIV	Recall: 99.8%	Employed a high-resolution CNN for feature extraction and disparity estimation.	Registration may fail in occluded or weak-texture regions; high-resolution features increase cost.
Tang et al. [3]	RGB-T object tracking	Attention-based multimodal fusion network	GTOT	Success rate: 68.7%	Utilizes pixel-level, attention-based feature-level, and dynamic decision-level fusion.	Multi-stage fusion increases computation and depends on synchronized RGB-T inputs.
Farahnakian et al. [41]	Maritime OD	CNN	Marine	AP: 79.1%	Uses multimodal fusion to detect marine vessels with high reliability.	Small vessels are difficult due to limited pixels, downsampling, and sea-condition variation.

Table 3. *Cont.*

Study	Task Type	Architecture	Datasets	Reported Metric/Task Outcome	Proposed Method	Limitations
Hu et al. [69]	VIR-OD	YOLOX	LLVIP	AP: 69	Uses light-sensing strategy to combine detections and enhance low-light performance.	Late decision fusion may miss early cross-modal cues; small objects remain difficult.
Thaket et al. [70]	Multispectral OD	Faster R-CNN with FPN	KAIST FLIR	mAP: 57.9 mAP: 78.9	Integrates DL-based multispectral detector in a feature pyramid network to improve low-light performance.	FPN fusion increases computation; RGB–thermal misalignment affects small/occluded objects.
Ying-Cheng et al. [71]	ADAS OD	Faster R-CNN	FLIR	EN: 7.06	Combines features via fusion for low-light or poor visibility conditions.	Depends on visible–IR alignment; fused images may lose fine small-object details.
Osin et al. [72]	Multispectral OD	CNN	KAIST	mAP: 69.45	Enhances SSD architecture with feature extraction and fusion for improved low-light performance.	Embedded deployment is limited by memory/speed trade-offs and possible accuracy loss.

Note: TNO and RoadScene are primarily image-fusion datasets; therefore, when they appear in Table 3, they should be understood as datasets used for fused-image quality evaluation rather than as standard OD benchmarks. Similarly, MS-COCO is an RGB-only dataset and is included only where it was used for visible-image pretraining, baseline development, or domain-adaptation experiments, not as a VIR dataset.

4.3. Studies on FER with Visible and IR Images

Compared with OD, VIR-FER remains less extensively studied. The limited number of available FER studies reflects the scarcity of paired VIR facial-expression datasets, the difficulty of collecting synchronized facial thermal and visible data, and the lack of standardized evaluation protocols. Existing VIR-FER studies provide useful evidence that thermal information can complement visible facial cues under illumination variation, partial visibility, and sensor-dependent imaging conditions; however, current evidence is still insufficient for broad cross-dataset conclusions. Therefore, FER is treated in this review as an emerging human-centered application of VIR fusion rather than as a domain with the same level of maturity as RGB-T OD.

The need for robust FER has motivated multimodal approaches that combine visible facial appearance with IR or thermal facial cues. Table 4 summarizes the available VIR-FER studies, including their datasets, reported metrics, fusion strategies, and limitations. Because the number of directly relevant VIR-FER studies is limited, the Table is interpreted as an evidence map of an emerging research direction rather than as a mature benchmark comparison. Early VIR-FER studies mainly investigated whether visible and IR facial cues could improve expression recognition compared with single-modality analysis. Siddiqui et al. [73] used visible and IR images from the VIRI dataset and proposed a CNN-based multimodal FER framework in which features from both modalities were combined and classified using an SVM to recognize five expressions. Although the method achieved 82.26% accuracy, its generalizability remains limited by the size and demographic distribution of the available dataset. Naseem et al. [74] further examined early- and late-fusion strategies using a modified CNN with 1- and 3-step training. Their results showed that concatenated visible and IR features achieved 84.44% accuracy on the VIRI dataset and 84% accuracy on the NVIE dataset, while additional experiments using the MSX modality demonstrated the potential of multispectral image-level fusion for improving FER robustness. Another attention-based early-fusion study further examined visible and IR facial-expression recognition by integrating modality-specific deep features with an attention mechanism, demonstrating the potential of attention-guided VIR fusion for FER [75]. Other VIR-FER studies show that both image-level preprocessing and feature-/decision-level integration can improve facial-expression classification, although their evaluation settings differ. The CNN-based method evaluated on the NIST dataset achieved high recognition accuracy by learning

discriminative facial features from thermal or multimodal facial data [76]. In contrast, the method evaluated on the NVIE dataset combined visible-spectrum facial information with thermal cues through a three-step framework based on Active Appearance Model features, head-motion features, statistical feature selection using the F-test, and classification with Bayesian networks and support vector machines [77]. This approach represents a combination of feature-level and decision-level fusion because selected visible and thermal features are integrated before classification, and modality-specific recognition outputs are further combined. However, the reported confusion between disgust and happiness in IR images indicates that thermal facial patterns alone may not always preserve expression-specific details, especially for visually subtle or texture-dependent expressions.

Tran et al. [78] proposed a decision-level fusion method that combined visible and IR facial images using ResNet-50-based transfer learning models trained separately on visible and thermal images. The final decision was obtained through an adaptive weighted fusion strategy based on the F1-scores of the modality-specific models. Using the KTFE dataset, the fusion method achieved an F1-score of 96.09%. Another method [46] used architectures such as CNNs, ResNet50, and YOLO to recognize multimodal expressions. The method was tested on the KTFEv2 dataset, using a CNN and a pre-trained ResNet50. Based on previous experiments, they combined manual feature extraction methods, achieving 86.8% accuracy. They also tested YOLOv5 and YOLOv7. The method divides major expressions (happy, sad, and angry) into three categories (low, medium, and high) to estimate intensities. Another study uses the IRIS dataset to identify features across multiple modalities [79]. The method used CNNs, autoencoders (AEs), and NNs to recognize three expressions. For the NN, accuracy was 93.3%, and for AE and CNN, accuracies of 90% and 96.7%, respectively, were achieved.

Table 4. Summary of multimodal VIR studies for FER.

Study	Architecture	Datasets	Reported Metric	Proposed Method	Limitations
Siddiqui et al. [73]	CNN and SVM	VIRI	Accuracy: 82.26%	Early fusion method using CNN and SVM	Limited dataset size/diversity; feature concatenation may overfit paired VIR samples.
Naseem et al. [74]	CNN	VIRI NVIE	Accuracy: 84.44%; Accuracy: 84%	Early and late fusion methods using 1- and 3-step training	IR features may lose subtle eye, eyebrow, and mouth texture; alignment is important.
Naseem et al. [75]	CNN with attention mechanism	VIRI NVIE	Accuracy: 84.44% Accuracy: 85.20%	Uses early fusion of visible and IR deep features with an attention mechanism for FER.	Limited to available paired VIR-FER data; performance depends on facial alignment and dataset diversity.
Mehendale [76]	CNN	NIST	Accuracy: 96%	Uses a two-part CNN: first removes background, second extracts facial features.	Sensitive to face/background extraction errors, shadows, occlusion, and low contrast.
Wang et al. [77]	Bayesian Network SVM	NVIE	Accuracy: 76.82%	The Bayesian network and SVM	Thermal cues may confuse similar expressions; depends on AAM extraction and alignment.
Tran et al. [78]	ResNet-50	KTFE	F1-score: 96.09%	Decision-level fusion using adaptive weighted fusion of visible and thermal ResNet-50 models.	Decision fusion depends on each ResNet-50 model and may miss subtle cross-modal cues.
Nguyen et al. [46]	CNN Resnet50 YOLO	KTFEv2	Accuracy: 86.8%	FER, along with intensity estimation, has been proposed using ML	Intensity estimation depends on selected frames and controlled acquisition settings.
Elbarawy et al. [79]	NN AE CNN	IRIS	Accuracy: 93.3%; 90%; 96.7%	FER has been proposed using NN, AE, and CNN	Limited subjects/classes; thermal-only features may miss texture-dependent expressions.

Note: The reported FER metrics are classification-based measures, mainly accuracy and F1-score. These values should be interpreted within the dataset, modality configuration, class distribution, train/test split, backbone model, and evaluation protocol used in each study. They are not intended for direct ranking against image-fusion quality metrics or OD/tracking metrics reported in other tables.

Overall, the reviewed FER studies indicate that VIR fusion can improve recognition robustness by combining visible texture and appearance information with illumination-invariant thermal cues. However, the current VIR-FER literature remains limited by small datasets, restricted demographic diversity, controlled acquisition settings, inconsistent emotion categories, and limited cross-dataset validation. These limitations make it difficult to draw strong general conclusions about real-world FER performance. Future VIR-FER studies should therefore prioritize larger paired VIR facial-expression datasets, standardized train/test protocols, demographic and environmental diversity, cross-dataset evaluation, and lightweight fusion architectures suitable for real-time human–computer interaction.

5. Cross-Study Synthesis of Fusion Strategies

Section 4 summarizes individual application-specific studies on VIR-OD and FER. To avoid repeating the same study-level information, this section provides a cross-study synthesis of fusion strategies. In this synthesis, fusion categories are assigned according to the stage at which visible and IR information is combined in the dataflow, rather than as mutually exclusive labels for entire papers. Therefore, multi-strategy studies may appear in more than one category when they perform fusion at multiple stages. For example, a method may first generate a fused image through sensor-/pixel-level fusion and later combine modality-specific features or decisions inside the detection or recognition pipeline. This criterion explains why studies such as [3,41,58] can be associated with more than one fusion level. The discussion focuses on what each fusion level contributes across tasks, what limitations repeatedly appear across studies, and how these limitations affect practical deployment in OD and FER. Therefore, Tables 5 and 6 are intended as synthesis Tables that highlight common strengths and weaknesses rather than duplicate summaries of the studies already listed in Tables 3 and 4.

Table 5. Comparative summary of fusion-level strategies for VIR OD.

Methods		Strengths	Weaknesses	Reference
Early fusion	Sensor-/pixel-level fusion	Produces fused images before detection and improves target visibility under low illumination, haze, smoke, and poor-visibility conditions. It is useful when interpretable fused images are required.	Highly sensitive to image registration errors, sensor-resolution mismatch, and fusion artifacts; fused-image quality does not always guarantee better detection accuracy.	[3,41,58]
Late fusion	Feature-level fusion	Combines modality-specific visible and IR representations inside detection networks, enabling task-adaptive fusion and improved robustness in low-light or complex scenes.	Requires paired multimodal data, careful network design, higher computational cost, and strong modality alignment.	[3,41,54–57,59–67,70,72]
	Score-level fusion	Combines confidence scores or prediction probabilities from modality-specific models, allowing late integration without modifying internal feature extractors.	Limited cross-modal interaction because fusion occurs only at the output-score stage; performance depends on reliable modality-specific models.	[72]
	Decision-level fusion	Combines final decisions from visible and IR models using rule-based, voting, or confidence-based strategies, providing modular and interpretable integration.	May miss fine-grained complementary information between modalities; performance is affected by errors made before the decision stage.	[3,41,69]

Note: Some studies are listed under more than one fusion category because they use more than one fusion stage. The categories refer to the fusion operation and dataflow stage used in the method, not to a mutually exclusive classification of each paper.

Table 6. Comparative summary of fusion-level strategies for VIR-FER.

Methods		Strengths	Weaknesses	References
Early fusion	Sensor-/pixel-level fusion	Combines VIR information at the image level, potentially improving facial detail, contrast, and illumination robustness before recognition.	Requires accurate facial alignment and synchronized VIR capture; limited validation on large and diverse FER datasets.	[73,74]
	Feature-level fusion	Integrates visible texture features and IR/thermal facial cues inside the recognition model, often improving recognition accuracy compared with single-modality inputs.	Sensitive to dataset size, demographic imbalance, controlled acquisition settings, and overfitting; cross-dataset validation remains limited.	[46,73–77,79]
Late fusion	Score-level fusion	Combines modality-specific confidence scores and allows flexible late integration of visible, infrared, or MSX-based models.	Depends strongly on the reliability of each modality-specific model; limited evidence across datasets and expression categories.	[74]
	Decision-level fusion	Combines final predictions from visible and IR models, providing a simple and modular multimodal FER framework.	Fusion occurs after individual predictions, so subtle complementary cues may be lost, affected by thermal noise, eyeglasses, and ambient temperature variation.	[46,77,78]

Note: The fusion categories refer to the stage at which visible and IR facial information is combined. Multi-stage FER methods may therefore be associated with more than one fusion level when image-level, feature-level, score-level, or decision-level information is combined within the same framework.

5.1. Fusion Methods for Object Detection

For OD, the reviewed studies show that different fusion levels support different stages of the detection pipeline. Sensor-/pixel-level fusion improves visual target saliency before detection, feature-level fusion enables task-adaptive representation learning inside detection networks, and score- or decision-level fusion provides modular integration of modality-specific outputs. Unlike Table 3, which summarizes individual studies, Table 5 synthesizes the recurring strengths and limitations of each fusion level across the reviewed OD literature.

Overall, feature-level fusion appears to be the most commonly used strategy in recent detection-oriented VIR studies because it allows the fusion process to be optimized together with detection objectives [54–67,70,72]. Sensor-/pixel-level fusion remains useful when a visually interpretable fused image is required, particularly in fusion-then-detect pipelines, but its dependence on accurate registration limits robustness in uncontrolled environments [3,41,58]. Score-level fusion and decision-level fusion provide simpler modular integration of modality-specific outputs, but they generally allow weaker interaction between visible and infrared cues than feature-level fusion [3,41,69,72]. Across all fusion levels, recurring challenges include multimodal alignment, dataset dependence, computational complexity, and limited validation under dynamic real-world conditions [54–63,66–72].

5.2. Fusion Methods for FER

For FER, the available VIR studies are fewer than those for OD, but they still show how different fusion levels can support human-centered affective analysis. Unlike Table 4, which summarizes individual VIR-FER studies, Table 6 synthesizes the main benefits and recurring limitations of each fusion level for FER.

Overall, the FER literature suggests that feature-level fusion is promising because it can combine visible facial texture with IR or thermal facial cues before classification [46,73–77,79]. However, the small number of paired VIR FER datasets limits the strength of current conclusions [15,44–47,73,74]. Sensor-/pixel-level fusion and MSX-style image-level fusion can improve visual interpretability, but they depend on accurate face alignment and synchronized acquisition [73,74]. Score-level and decision-level fusion provide simpler late-fusion alternatives, but they may not fully exploit subtle cross-modal relationships between visible appearance and thermal facial information [46,74,77,78]. Therefore, future VIR-FER research

should focus on larger paired datasets, standardized protocols, cross-dataset validation, and lightweight fusion networks suitable for real-time human–computer interaction [46,73,74,78].

6. OD- and FER-Oriented Applications

VIR fusion has broad application potential, but in the context of this review, its practical relevance is mainly considered through OD and FER. For OD, VIR fusion improves target localization and recognition under low illumination, adverse weather, smoke, haze, occlusion, and complex backgrounds [80]. For FER, VIR fusion provides complementary thermal and illumination-invariant facial cues that may improve affective analysis when visible facial texture is unreliable [81]. Therefore, the following applications are discussed according to how they benefit OD- and FER-oriented computer vision systems rather than as general image-fusion applications.

6.1. Human Emotion Identification

Human emotion identification is the most direct FER-oriented application of VIR fusion. Visible facial images provide texture, shape, and appearance cues, whereas IR or thermal images can capture illumination-invariant facial information and physiological variations related to facial temperature distribution [82]. This combination is useful when visible images are affected by low illumination, shadows, or partial visibility, and when thermal facial information complements visible facial details [45]. However, current VIR-FER applications remain limited by small paired datasets, demographic imbalance, controlled acquisition settings, and sensitivity of thermal images to eyeglasses, ambient temperature, and sensor quality. Therefore, practical FER deployment requires larger visible–IR facial-expression datasets, stronger cross-dataset validation, and privacy-aware human–computer interaction frameworks.

6.2. Surveillance

Surveillance is mainly connected to the OD focus of this review. In nighttime, smoke, haze, fog, and low-visibility environments, visible cameras may fail to capture reliable texture and color information, whereas IR images can highlight pedestrians, vehicles, animals, and other heat-emitting targets [83]. VIR fusion can therefore improve detection robustness for security monitoring, search-and-rescue operations, traffic monitoring, and perimeter protection [84]. For FER, surveillance-oriented use is more sensitive and should be treated cautiously because facial emotion analysis in public or semi-public environments raises privacy, consent, demographic-bias, and ethical concerns. Thus, OD is the more mature surveillance application, while FER requires stricter ethical and technical safeguards before deployment.

6.3. Medical

Medical applications are relevant to this review mainly when VIR fusion supports detection or recognition tasks, such as identifying abnormal thermal patterns, monitoring facial affect in healthcare interaction, or supporting assistive human–computer interfaces. The fusion process combines detailed anatomical information from visible-light images with functional data from IR imaging, such as blood flow or metabolic heat patterns [85]. This integrated approach improves diagnosis and monitoring of medical conditions, offering clinicians a comprehensive view of physiological processes and abnormalities [86]. Fused images can help detect skin cancers early by highlighting abnormal thermal signatures alongside visible dermatological features, or aid in managing vascular diseases by combining visual and thermal images to assess blood flow and tissue health.

6.4. Remote Sensing

Remote sensing is relevant to the OD focus of this review because VIR fusion can improve the detection of vehicles, vessels, fires, vegetation stress, and other targets under complex environmental conditions. This technology enhances remote sensing platforms' ability to monitor environmental conditions, detect vegetation changes, and manage natural disasters [87]. By integrating visible-light images with IR data, applications can achieve a better understanding of environmental phenomena [88]. Fused images can improve forest fire detection by combining visible imagery with IR heat signals, and help assess drought impact by correlating temperatures with vegetation health. The applications of visible and IR image fusion span multiple fields, benefiting from enhanced interpretation capabilities [26,89]. While these surveys provide broad coverage of remote sensing and infrared–visible image-fusion applications, the present review narrows the discussion to OD- and FER-oriented computer vision tasks and emphasizes dataset suitability, task-specific metrics, fusion-level design, and deployment limitations.

7. Future Directions and Challenges

Future research on VIR fusion should move beyond general fused-image enhancement and focus more directly on downstream OD and FER performance. For OD, key challenges include robust multimodal alignment, small-object detection, real-time inference, adverse-weather generalization, and evaluation on diverse VIR benchmarks. For FER, the major challenges are the scarcity of paired VIR expression datasets, limited demographic diversity, inconsistent emotion labels, thermal sensitivity to external conditions, and weak cross-dataset validation. Addressing these issues is essential for developing reliable VIR-based computer vision systems.

7.1. Enhancing OD and FER Accuracy

Improving OD and FER accuracy requires fusion methods that remain reliable under sensor misalignment, illumination variation, occlusion, adverse weather, and dynamic background conditions [90,91]. For OD, future models should integrate alignment-robust fusion modules, task-aware losses, and cross-modal attention mechanisms that directly improve detection performance rather than only fused-image quality. For FER, accuracy improvement requires larger and more diverse VIR facial-expression datasets, better handling of thermal artifacts, and validation across subjects, environments, sensors, and demographic groups [15,44–47,73,74]. Evaluation should therefore include both in-dataset and cross-dataset testing to assess real-world generalization [92].

7.2. Achieving Real-Time OD and FER Processing

Real-time OD and FER processing remains challenging because multimodal fusion increases computational cost and latency [93]. For OD, lightweight detectors, efficient feature-fusion modules, pruning, quantization, and edge-computing deployment are needed for surveillance, traffic monitoring, autonomous systems, and embedded platforms [94]. For FER, real-time deployment requires compact fusion networks that can process visible and IR facial inputs without excessive delay, while maintaining robustness to head pose, facial occlusion, eyeglasses, and thermal noise. Future systems should report latency, model size, and hardware conditions in addition to accuracy or mAP.

7.3. Standardized Evaluation and Dataset Expansion

Standardized evaluation and dataset expansion are essential for both OD and FER. For OD, datasets such as LLVIP, M3FD, DroneVehicle, FLIR, and KAIST provide useful VIR or RGB–thermal benchmarks, but studies still differ in task setting, annotation

quality, modality alignment, and reported metrics [31–33,40,42]. Future OD studies should use consistent train/test protocols, report AP or mAP with clear modality settings, and evaluate robustness under low light, weather variation, occlusion, and cross-dataset transfer. For FER, the need is more urgent because available paired VIR facial-expression datasets remain limited in size, demographic diversity, and emotion-label consistency [15,44–47]. Future VIR-FER datasets should include diverse subjects, controlled and in-the-wild conditions, standardized emotion labels, synchronized VIR acquisition, and public evaluation protocols.

7.4. Ethical and Deployment Considerations

Ethical and deployment considerations are particularly important when VIR fusion is used for human-centered OD and FER systems. OD applications in surveillance, traffic monitoring, and smart transportation can benefit from improved nighttime and adverse-weather detection, but deployment must consider reliability, false alarms, sensor cost, and performance under real-world domain shifts [95,96]. FER applications require even greater caution because facial emotion analysis can involve sensitive personal information, demographic bias, privacy risks, and possible misuse in public or workplace monitoring. Future FER systems should therefore incorporate privacy-aware data collection, consent procedures, bias assessment, transparent reporting, and clear limits on deployment contexts.

8. Conclusions

We reviewed multimodal VIR image-fusion techniques for computer vision applications, with OD and FER considered as two representative downstream tasks. By focusing on these applications, this review demonstrates how VIR fusion supports both scene-level perception and human-centered affective analysis under challenging imaging conditions. Our analysis of state-of-the-art methods and datasets revealed limitations in current techniques, indicating areas for improvement. We examined early fusion (sensor-level and feature-level) and late fusion (score-level and decision-level) methods, comparing their effectiveness for designing advanced models. OD and FER are important for surveillance, security, healthcare, and human–computer interaction, where accurate scene perception and emotional analysis are essential. The integration of modalities shows potential across surveillance, medical imaging, and remote sensing. This review demonstrates the impact of VIR fusion while highlighting the need for further research. Future work should focus on developing robust fusion techniques for dynamic environments and real-time processing, along with exploring novel datasets, advanced models, and AI-based techniques to enhance multimodal detection systems.

Author Contributions: M.T.N.: Conceptualization, methodology, data curation, formal analysis, writing—review and editing, and visualization. C.-S.L.: Supervision, conceptualization, project administration, funding acquisition, and writing—review and editing. M.A.K.: Investigation, editing, resources, and data analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Technology Innovation Program (20019078) funded by the Ministry of Trade, Industry & Energy (MOTIE, Republic of Korea) and by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2021-NR060125).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

CNN	Convolutional neural network
DL	Deep learning
ML	Machine learning
FE	Facial expression
FER	Facial-expression recognition
GE-WA	Gaussian estimation-weighted average
IR	Infrared
Vis	Visible
NIR	Near-infrared
TIR	Thermal infrared
TNO	The Netherlands Organisation for Applied Scientific Research
VIR	Visible–infrared
SOTA	State-of-the-art

References

1. Liu, J.; Liu, Z.; Wu, G.; Ma, L.; Liu, R.; Zhong, W.; Luo, Z.; Fan, X. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023.
2. Yin, W.; He, K.; Xu, D.; Yue, Y.; Luo, Y. Adaptive low light visual enhancement and high-significant target detection for infrared and visible image fusion. *Vis. Comput.* **2023**, *39*, 6723–6742. [CrossRef]
3. Tang, Z.; Xu, T.; Li, H.; Wu, X.-J.; Zhu, X.; Kittler, J. Exploring fusion strategies for accurate RGBT visual object tracking. *Inf. Fusion* **2023**, *99*, 101881. [CrossRef]
4. Farahnakian, F.; Movahedi, P.; Poikonen, J.; Lehtonen, E.; Makris, D.; Heikkonen, J. Comparative analysis of image fusion methods in marine environment. In *Proceedings of the 2019 IEEE International Symposium on Robotic and Sensors Environments (ROSE), Ottawa, ON, Canada, 17–18 June 2019*; IEEE: New York, NY, USA, 2019.
5. Haghbayan, M.-H.; Farahnakian, F.; Poikonen, J.; Laurinen, M.; Nevalainen, P.; Plosila, J. An efficient multi- sensor fusion approach for object detection in maritime environments. In *Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018*; IEEE: New York, NY, USA, 2018.
6. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
7. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
8. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv* **2013**, arXiv:1312.6229.
9. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
10. Li, H.; Wu, X.-J. DenseFuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.* **2019**, *28*, 2614–2623. [CrossRef]
11. Gao, H.; Cheng, B.; Wang, J.; Li, K.; Zhao, J.; Li, D. Object classification using CNN-based fusion of vision and LIDAR in autonomous vehicle environment. *IEEE Trans. Ind. Inform.* **2018**, *14*, 4224–4231. [CrossRef]
12. Xue, S.; Liu, Y.; Xu, C.; Li, J. Object detection in visible and infrared missile borne fusion image. In *Proceedings of the 2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), Xi'an, China, 28–30 October 2022*; IEEE: New York, NY, USA, 2022.
13. Krishnan, B.S.; Jones, L.R.; Elmore, J.A.; Samiappan, S.; Evans, K.O.; Pfeiffer, M.B.; Blackwell, B.F.; Iglay, R.B. Fusion of visible and thermal images improves automated detection and classification of animals for drone surveys. *Sci. Rep.* **2023**, *13*, 10385. [CrossRef] [PubMed]

14. Fan, W.; Li, X.; Liu, Z. Fusion of visible and infrared images using GE-WA model and VGG-19 network. *Sci. Rep.* **2023**, *13*, 190. [CrossRef] [PubMed]
15. Siddiqui, I.F.H.; Dhakal, P.; Yang, X.; Javaid, A.Y. A survey on databases for multimodal emotion recognition and an introduction to the VIRI (visible and InfraRed image) database. *Multimodal Technol. Interact.* **2022**, *6*, 47. [CrossRef]
16. Jaimes, A.; Sebe, N. Multimodal human–computer interaction: A survey. *Comput. Vis. Image Underst.* **2007**, *108*, 116–134. [CrossRef]
17. Alonso-Martin, F.; Malfaz, M.; Sequeira, J.; Gorostiza, J.F.; Salichs, M.A. A multimodal emotion detection system during human–robot interaction. *Sensors* **2013**, *13*, 15549–15581. [CrossRef] [PubMed]
18. Busso, C.; Deng, Z.; Yildirim, S.; Bulut, M.; Lee, C.M.; Kazemzadeh, A.; Lee, S.; Neumann, U.; Narayanan, S. Analysis of emotion recognition using facial expressions, speech and multimodal information. In Proceedings of the 6th International Conference on Multimodal Interfaces, New York, NY, USA, 13–15 October 2004.
19. De Silva, L.C.; Miyasato, T.; Nakatsu, R. Facial emotion recognition using multi-modal information. In *Proceedings of the ICICS, 1997 International Conference on Information, Communications and Signal Processing, Singapore, 9–12 September 1997*; IEEE: New York, NY, USA, 1997.
20. Wagner, J.; Andre, E.; Lingenfelser, F.; Kim, J. Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Trans. Affect. Comput.* **2011**, *2*, 206–218. [CrossRef]
21. Huang, Y.; Yang, J.; Liao, P.; Pan, J. Fusion of facial expressions and EEG for multimodal emotion recognition. *Comput. Intell. Neurosci.* **2017**, *2017*, 2107451. [CrossRef] [PubMed]
22. Roseman, J. Emotional behaviors, emotivational goals, emotion strategies: Multiple levels of organization integrate variable and consistent responses. *Emot. Rev.* **2011**, *3*, 434–443. [CrossRef]
23. Wen, W.; Liu, G.; Cheng, N.; Wei, J.; Shangguan, P.; Huang, W. Emotion recognition based on multi-variant correlation of physiological signals. *IEEE Trans. Affect. Comput.* **2014**, *5*, 126–140. [CrossRef]
24. Jiang, Y.; Li, W.; Hossain, M.S.; Chen, M.; Alelaiwi, A.; Al-Hammadi, M. A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Inf. Fusion* **2020**, *53*, 209–221. [CrossRef]
25. Yang, K.; Xiang, W.; Chen, Z.; Zhang, J.; Liu, Y. A review on infrared and visible image fusion algorithms based on neural networks. *J. Vis. Commun. Image Represent.* **2024**, *101*, 104179. [CrossRef]
26. Ma, J.; Ma, Y.; Li, C. Infrared and visible image fusion methods and applications: A survey. *Inf. Fusion* **2019**, *45*, 153–178. [CrossRef]
27. Toet, A. The TNO multiband image data collection. *Data Brief* **2017**, *15*, 249. [CrossRef] [PubMed]
28. Xu, H.; Ma, J.; Jiang, J.; Guo, X.; Ling, H. U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 502–518. [CrossRef] [PubMed]
29. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference Proceedings, Part V, Zurich, Switzerland, 6–12 September 2014.
30. Tang, L.; Yuan, J.; Zhang, H.; Jiang, X.; Ma, J. PIAFusion: A Progressive Infrared and Visible Image Fusion Network Based on Illumination Aware. *Inf. Fusion* **2022**, *83*, 79–92. [CrossRef]
31. Jia, X.; Zhu, C.; Li, M.; Tang, W.; Zhou, W. LLVIP: A visible-infrared paired dataset for low-light vision. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
32. M3FD: Multi-Scenario Multi-Modality Dataset for Infrared and Visible Image Fusion and Object Detection. Available online: https://github.com/JinyuanLiu-CV/TarDAL (accessed on 21 May 2026).
33. Sun, Y.; Cao, B.; Zhu, P.; Hu, Q. Drone-Based RGB-Infrared Cross-Modality Vehicle Detection via Uncertainty-Aware Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6700–6713. [CrossRef]
34. Li, C.; Liang, X.; Lu, Y.; Zhao, N.; Tang, J. Weighted sparse representation regularized graph learning for RGB-T object tracking. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017.
35. Li, C.; Liang, X.; Lu, Y.; Zhao, N.; Tang, J. RGB-T object tracking: Benchmark and baseline. *Pattern Recognit.* **2019**, *96*, 106977. [CrossRef]
36. Davis, I.W.; Sharma, V. Background-subtraction using contour-based fusion of thermal and visible imagery. *Comput. Vis. Image Underst.* **2007**, *106*, 162–182. [CrossRef]
37. Torabi, A.; Massé, G.; Bilodeau, G.-A. An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications. *Comput. Vis. Image Underst.* **2012**, *116*, 210–221. [CrossRef]
38. Li, A.; Cheng, H.; Hu, S.; Liu, X.; Tang, J.; Lin, L. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Trans. Image Process.* **2016**, *25*, 5743–5756. [CrossRef]
39. INO. Video Analytics Dataset. Available online: https://www.ino.ca/en/technologies/video-analytics-dataset (accessed on 21 May 2026).

40. Farooq, M.A.; Corcoran, P.; Rotariu, C.; Shariff, W. Object detection in thermal spectrum for advanced driver-assistance systems (ADAS). *IEEE Access* **2021**, *9*, 156465–156481. [CrossRef]
41. Farahnakian, F.; Heikkonen, J. Deep learning based multi-modal fusion architectures for maritime vessel detection. *Remote Sens.* **2020**, *12*, 2509. [CrossRef]
42. Hwang, S.; Park, J.; Kim, N.; Choi, Y.; Kweon, I.S. Multispectral pedestrian detection: Benchmark dataset and baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
43. Agarwal, S.; Sikchi, H.S.; Rooj, S.; Bhattacharya, S.; Routray, A. Illumination-invariant face recognition by fusing thermal and visual images via gradient transfer. In *Advances in Computer Vision, Proceedings of the 2019 Computer Vision Conference (CVC)*; Springer: Cham, Switzerland, 2020; Volume 1.
44. Wang, S.; Liu, Z.; Lv, S.; Lv, Y.; Wu, G.; Peng, P.; Chen, F.; Wang, X. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Trans. Multimed.* **2010**, *12*, 682–691. [CrossRef]
45. Nguyen, H.; Kotani, K.; Chen, F.; Le, B. A thermal facial emotion database and its analysis. In *Image and Video Technology, Proceeding of the Image and Video Technology: 6th Pacific-Rim Symposium, PSIVT 2013, Guanajuato, Mexico, 28 October–1 November 2013*; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8333.
46. Nguyen, H.; Tran, N.; Nguyen, H.D.; Nguyen, L.; Kotani, K. KTFEv2: Multimodal facial emotion database and its analysis. *IEEE Access* **2023**, *11*, 17811–17822. [CrossRef]
47. Kowalski, I.; Grudzień, A. High-resolution thermal face dataset for face and expression recognition. *Metrol. Meas. Syst.* **2018**, *25*, 403–415. [CrossRef]
48. Pereira, L.M.; Salazar, A.; Vergara, L. A comparative study on recent automatic data fusion methods. *Computers* **2024**, *13*, 2006. [CrossRef]
49. Roberts, J.W.; Van Aardt, J.A.; Ahmed, F.B. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *J. Appl. Remote Sens.* **2008**, *2*, 023522. [CrossRef]
50. Wang, P.-w.; Liu, B. A novel image fusion metric based on multi-scale analysis. In *Proceeding of the 2008 9th International Conference on Signal Processing, Beijing, China, 26–29 October 2008*; IEEE: New York, NY, USA, 2008.
51. James, A.P.; Dasarathy, B.V. Medical image fusion: A survey of the state of the art. *Inf. Fusion* **2014**, *19*, 4–19. [CrossRef]
52. Han, Y.; Cai, Y.; Cao, Y.; Xu, X. A new image fusion performance metric based on visual information fidelity. *Inf. Fusion* **2013**, *14*, 127–135. [CrossRef]
53. Yuan, Y.; Zhang, J.; Chang, B.; Han, Y. Objective quality evaluation of visible and infrared color fusion image. *Opt. Eng.* **2011**, *50*, 033202. [CrossRef]
54. Liu, Y.; Zhou, X.; Zhong, W. Multi-modality image fusion and object detection based on semantic information. *Entropy* **2023**, *25*, 718. [CrossRef] [PubMed]
55. 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 Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022.
56. Tu, X.; Yuan, Z.; Liu, B.; Liu, J.; Hu, Y.; Hua, H.; Wei, L. An improved YOLOv5 for object detection in visible and thermal infrared images based on contrastive learning. *Front. Phys.* **2023**, *11*, 1193245. [CrossRef]
57. Kim, J.; Huh, J.; Park, I.; Bak, J.; Kim, D.; Lee, S. Small object detection in infrared images: Learning from imbalanced cross-domain data via domain adaptation. *Appl. Sci.* **2022**, *12*, 11201. [CrossRef]
58. Zhang, X.; Zhai, H.; Liu, J.; Wang, Z.; Sun, H. Real-time infrared and visible image fusion network using adaptive pixel weighting strategy. *Inf. Fusion* **2023**, *99*, 101863. [CrossRef]
59. Yang, F.; Cheng, I. Visible-infrared features fusion based object detection. In *Proceeding of the 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Honolulu, Oahu, HI, USA, 1–4 October 2023*; IEEE: New York, NY, USA, 2023.
60. Zhao, T.; Yuan, M.; Jiang, F.; Wang, N.; Wei, X. Removal and selection: Improving RGB-infrared object detection via coarse-to-fine fusion. *arXiv* **2024**, arXiv:2401.10731.
61. Jiang, C.; Ren, H.; Yang, H.; Huo, H.; Zhu, P.; Yao, Z.; Li, J.; Sun, M.; Yang, S. M^2FNet: Multi-modal fusion network for object detection from visible and thermal infrared images. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *130*, 103918. [CrossRef]
62. Deevi, S.A.; Lee, C.; Gan, L.; Nagesh, S.; Pandey, G.; Chung, S.-J. RGB-X object detection via scene-specific fusion modules. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2024; pp. 7366–7375.
63. Hou, Z.; Yang, C.; Sun, Y.; Ma, S.; Yang, X.; Fan, J. An object detection algorithm based on infrared-visible dual modal feature fusion. *Infrared Phys. Technol.* **2024**, *137*, 105107. [CrossRef]
64. Meng, F.; Hong, A.; Tang, H.; Tong, G. FQDNet: A fusion-enhanced quad-head network for RGB-infrared object detection. *Remote Sens.* **2025**, *17*, 1095. [CrossRef]
65. Tan, W.; Geng, B.; Bai, X. A study on infrared-visible fusion multimodal object detection algorithm based on cross-modal information bottleneck and minimum redundancy transformation. *Sci. Rep.* **2026**, *16*, 12991. [CrossRef] [PubMed]

66. Shen, L.; Lang, B.; Song, Z. Infrared object detection method based on DBD-YOLOv8. *IEEE Access* **2023**, *11*, 145853–145868. [CrossRef]
67. Peng, Y.; Liu, G.; Xu, X.; Bavirisetti, D.P.; Gu, X.; Zhang, X. MFDetection: A highly generalized object detection network unified with multilevel heterogeneous image fusion. *Optik* **2022**, *266*, 169599. [CrossRef]
68. Li, Q.; Han, G.; Liu, P.; Yang, H.; Luo, H.; Wu, J. An infrared-visible image registration method based on the constrained point feature. *Sensors* **2021**, *21*, 1188. [CrossRef] [PubMed]
69. Hu, Z.; Jing, Y.; Wu, G. Decision-level fusion detection method of visible and infrared images under low light conditions. *EURASIP J. Adv. Signal Process.* **2023**, *2023*, 38.
70. Thaker, K.; Chennupati, S.; Rawashdeh, N.; Rawashdeh, S.A. Multispectral deep neural network fusion method for low-light object detection. *J. Imaging* **2023**, *10*, 12. [CrossRef] [PubMed]
71. Lin, Y.-C.; Chiang, P.-Y.; Miaou, S.-G. Enhancing deep-learning object detection performance based on fusion of infrared and visible images in advanced driver assistance systems. *IEEE Access* **2022**, *10*, 105214–105231. [CrossRef]
72. Osin, V.; Cichocki, A.; Burnaev, E. Fast multispectral deep fusion networks. *Bull. Pol. Acad. Sci. Tech. Sci.* **2018**, *66*, 875–889. [CrossRef]
73. Siddiqui, I.F.H.; Javaid, A.Y. A multimodal facial emotion recognition framework through the fusion of speech with visible and infrared images. *Multimodal Technol. Interact.* **2020**, *4*, 46. [CrossRef]
74. Naseem, M.T.; Lee, C.-S.; Kim, N.-H. Facial expression recognition using visible, IR, and MSX images by early and late fusion of deep learning models. *IEEE Access* **2024**, *12*, 20692–20704. [CrossRef]
75. Naseem, M.T.; Lee, C.S.; Shahzad, T.; Khan, M.A.; Abu-Mahfouz, A.M.; Ouahada, K. Facial expression recognition using visible and IR by early fusion of deep learning with attention mechanism. *PeerJ Comput. Sci.* **2025**, *11*, e2676. [CrossRef] [PubMed]
76. Mehendale, I. Facial emotion recognition using convolutional neural networks (FERC). *SN Appl. Sci.* **2020**, *2*, 446. [CrossRef]
77. Wang, S.; He, S.; Wu, Y.; He, M.; Ji, Q. Fusion of visible and thermal images for facial expression recognition. *Front. Comput. Sci.* **2014**, *8*, 232–242. [CrossRef]
78. Tran, K.; Nguyen, D.; Nguyen, D.; Tran, N.; Nguyen, H. Decision fusion from visible and infrared images for human emotion detection. *Int. J. Adv. Eng.* **2020**, *3*, 8–11.
79. Elbarawy, Y.M.; Ghali, N.I.; El-Sayed, R.S. Facial expressions recognition in thermal images based on deep learning techniques. *Int. J. Image Graph. Signal Process.* **2019**, *11*, 1–7. [CrossRef]
80. Pajares, G.; De La Cruz, J.M. A wavelet-based image fusion tutorial. *Pattern Recognit.* **2004**, *37*, 1855–1872. [CrossRef]
81. Mao, X.; Li, W.; Lei, C.; Jin, J.; Duan, F.; Chen, S. A brain–robot interaction system by fusing human and machine intelligence. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2019**, *27*, 533–542. [CrossRef] [PubMed]
82. Raheel, A.; Majid, M.; Alnowami, M.; Anwar, S.M. Physiological sensors based emotion recognition while experiencing tactile enhanced multimedia. *Sensors* **2020**, *20*, 4037. [CrossRef] [PubMed]
83. Ma, J.; Yu, W.; Liang, P.; Li, C.; Jiang, J. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **2019**, *48*, 11–26. [CrossRef]
84. Bebis, G.; Gyaourova, A.; Singh, S.; Pavlidis, I. Face recognition by fusing thermal infrared and visible imagery. *Image Vis. Comput.* **2006**, *24*, 727–742. [CrossRef]
85. De Souza, M.; Bueno, A.; Magas, V.; Neto, G.N.; Nohama, P. Imaging fusion between anatomical and infrared thermography of the thyroid gland. In *Proceedings of the 2019 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE), Buenos Aires, Argentina, 26–31 March 2019*; IEEE: New York, NY, USA, 2019.
86. Arpaia, I.; Manna, C.; Montenero, G.; D'Addio, G. In-time prognosis based on swarm intelligence for home-care monitoring: A case study on pulmonary disease. *IEEE Sens. J.* **2012**, *12*, 692–698. [CrossRef]
87. Ghassemian, H. A review of remote sensing image fusion methods. *Inf. Fusion* **2016**, *32*, 75–89. [CrossRef]
88. Liu, Z.; Yin, H.; Fang, B.; Chai, Y. A novel fusion scheme for visible and infrared images based on compressive sensing. *Opt. Commun.* **2015**, *335*, 168–177. [CrossRef]
89. Rattá, G.; Vega, J.; Murari, A.; Contributors, J. Image fusion: Advances in the state of the art. *Inf. Fusion* **2007**, *8*, 114–118. [CrossRef]
90. Irani, M.; Peleg, S. Motion analysis for image enhancement: Resolution, occlusion, and transparency. *J. Vis. Commun. Image Represent.* **1993**, *4*, 324–335. [CrossRef]
91. Fasbender, A.; Tuia, D.; Bogaert, P.; Kanevski, M. Support-based implementation of Bayesian data fusion for spatial enhancement: Applications to ASTER thermal images. *IEEE Geosci. Remote Sens. Lett.* **2008**, *5*, 598–602. [CrossRef]
92. Swets, J.A. Measuring the accuracy of diagnostic systems. *Science* **1988**, *240*, 1285–1293. [CrossRef] [PubMed]
93. Brooks, I.R.; Iyengar, S.S. Real-time distributed sensor fusion for time-critical sensor readings. *Opt. Eng.* **1997**, *36*, 767–779.
94. Ma, J.; Liang, P.; Yu, W.; Chen, C.; Guo, X.; Wu, J. Infrared and visible image fusion via detail preserving adversarial learning. *Inf. Fusion* **2020**, *54*, 85–98. [CrossRef]

95. Maimaitijiang, M.; Sagan, V.; Sidike, P.; Daloye, A.M.; Erkbol, H.; Fritschi, F.B. Crop monitoring using satellite/UAV data fusion and machine learning. *Remote Sens.* **2020**, *12*, 1357. [CrossRef]
96. Tang, J.; Ye, C.; Zhou, X.; Xu, L. YOLO-Fusion and Internet of Things: Advancing object detection in smart transportation. *Alex. Eng. J.* **2024**, *107*, 1–12. [CrossRef]