End-To-End Multi-View Multi-Modal Detection-Driven Image Fusion: One Method to Fuse them all

Gwendal Bernardi[1,2], Godefroy Brisebarre[1], Sébastien Roman[1], Mohsen Ardabilian[2], and Emmanuel Dellandrea[2]

[1] Tiama, Saint-Genis-Laval, France
[2] École Centrale de Lyon, CNRS, Université Claude Bernard Lyon 1, INSA Lyon, Université Lumière Lyon 2, LIRIS, UMR5205, 69130 Écully, France
g.bernardi@tiama.com, gwendal.bernardi@ec-lyon.fr

Abstract. We present EDIF, an end-to-end detection-driven framework designed to unify multi-modal and multi-view image fusion within a single architecture. While most existing fusion methods address either spectral complementarity (multi-modal) or viewpoint variability (multi-view) in isolation, real-world perception systems increasingly require both. EDIF formulates fusion as an object-level alignment problem: heterogeneous images are encoded as sets of keypoints, which are matched and aggregated through a graph attention mechanism to form object-centric representations directly optimized for detection. To stabilize training across heterogeneous components, we introduce a three-stage task-driven strategy that progressively aligns keypoint extraction, object localization, and cross-sensor grouping. In addition, we release the Multi-Modal and Multi-View Object Detection Dataset (MMDOD), a new benchmark designed to study detection-driven fusion under strong modality–view dependencies. MMDOD contains over 10,000 images of transparent objects captured under four complementary modalities (visible, NIR, low-contrast, polarization shift) and six viewpoints, with detailed object-level annotations. Experiments on RGB–thermal, multi-camera, and joint multi-modal multi-view benchmarks show that EDIF achieves performance competitive with recent specialized methods, while uniquely operating within a unified framework. On MMDOD, EDIF significantly outperforms adapted multi-modal multi-view baselines, highlighting the benefits of detection-driven, object-level fusion. The proposed MMDOD dataset is available at https://datasets.liris.cnrs.fr/mmdod-version1.

Keywords: Multi-View · Multi-Modal · Object Detection · Image Fusion · GNN

1 Introduction

Image fusion is a central problem in computer vision, enabling robust perception in applications ranging from medical imaging to autonomous driving [1] [19]. Beyond these domains, industrial environments increasingly rely on multi-modal

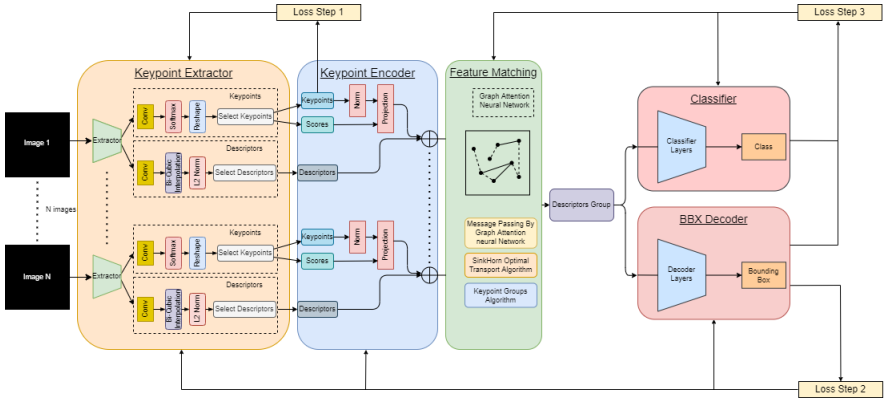

Fig. 1. Overview of the proposed EDIF pipeline. The architecture consists of four modules: (1) a keypoint extractor that identifies interest points in each image, (2) a keypoint encoder that projects descriptors into a shared representation, (3) a graph attention-based feature matching block that establishes correspondences across views and modalities, and (4) a decoder with two branches for bounding box regression and object classification. Training proceeds in three stages: keypoint localization with supervision from ground-truth box centers, bounding box regression and confidence estimation, and joint training of feature matching and classification with full detection supervision.

and multi-view perception systems for tasks such as automated quality control, robotic manipulation, defect detection, transparent object inspection, waste sorting, and production-line monitoring. In such contexts, a single viewpoint or sensing modality is often insufficient due to occlusions, specular reflections, transparency, low-texture surfaces, or harsh illumination conditions. Combining complementary observations from multiple sensors and viewpoints therefore becomes critical to ensure reliable and accurate perception in real-world industrial pipelines. By integrating complementary cues from heterogeneous sources, fusion improves robustness under challenging conditions such as occlusion, low illumination, or sensor degradation. Existing methods are typically specialized: multi-view fusion enforces geometric consistency across cameras, while multi-modal fusion integrates spectral or sensor-specific cues [2]. Real-world perception systems, however, frequently combine both dimensions simultaneously. This is particularly evident in industrial inspection setups involving multiple RGB, thermal, polarization, or X-ray sensors distributed across several viewpoints, as well as in multi-modal multi-view medical imaging [34] and autonomous driving platforms with multiple sensors and viewpoints [38] [35]. Despite this growing need, unified architectures capable of jointly handling multi-modal and multi-view fusion remain largely underexplored. Fusion approaches can be categorized into supervised, unsupervised, and task-driven paradigms [2]. Supervised methods rely on ground-truth fused images, which are often unavailable, while unsupervised strategies depend on heuristics weakly correlated with downstream tasks.

In industrial applications, these limitations are particularly problematic because perception systems must generalize robustly to unseen operating conditions while remaining directly optimized for actionable objectives such as detection, localization, or inspection. Task-driven learning, in contrast, supervises fusion through an end objective such as object detection, providing semantically grounded guidance. This motivates our detection-driven formulation of unified fusion. In this work, we introduce End-To-End Multi-View Multi-Modal Detection-Driven Image Fusion (EDIF), a generalist framework that unifies multi-modal and multi-view fusion within a single detection-oriented architecture. EDIF departs from dense feature aggregation strategies (e.g., MVMM-Net [34], GMMVF [6]) by formulating fusion as an object-level alignment problem. Input images are encoded as keypoints enriched with local descriptors, which are matched and refined through a graph attention network to reason jointly over spectral and viewpoint variations. Keypoints are then softly grouped into object-centric representations, optimized directly for bounding box regression and classification through a unified detection head. Stable end-to-end optimization is achieved via a three-stage training strategy that progressively aligns keypoint extraction, object localization, and cross-sensor grouping. Compared to pixel- or feature-level fusion, this object-level formulation enforces semantic alignment across heterogeneous inputs and mitigates the effect of appearance discrepancies caused by viewpoint or modality changes. This leads to more robust representations that are directly optimized for detection rather than reconstruction or low-level similarity.
To evaluate EDIF, we introduce the Multi-Modal and Multi-View Object Detection Dataset (MMDOD), a benchmark explicitly designed to study detection-driven fusion under strong modality-view dependencies. MMDOD contains over 10,000 annotated images of transparent industrial objects captured under four complementary modalities and six viewpoints, covering both destructive and non-destructive artifacts. The dataset specifically targets realistic industrial inspection challenges where appearance varies significantly across viewpoints and sensing technologies. For fair comparison, we adapt MVMM-Net into a detection-oriented baseline (MVMM-DET). MVMM-Net was originally designed for dense feature fusion and reconstruction tasks without detection supervision. We adapt it into MVMM-DET by introducing a detection head and task-driven training, while preserving its original multi-stage fusion design. Unlike EDIF, MVMM-DET performs modality and view fusion separately, without explicit object-level alignment.

Our main contributions are summarized as follows:

- An end-to-end detection-driven fusion framework. We propose EDIF, the first architecture capable of performing unified multi-modal and multi-view fusion for object detection, bridging both domains within a single, task-driven pipeline.
- A transformer-augmented keypoint extractor (TransPoint). We enhance SuperPoint with transformer layers for improved multi-modal and multi-view keypoint reliability.

- A dual-head detection and grouping strategy. EDIF introduces a GNN-based grouping mechanism to aggregate keypoints into object clusters, followed by dual-head classification and bounding-box prediction.
- A three-stage progressive learning paradigm. Our training schedule aligns keypoint, detection, and semantic grouping optimization with a decreasing σ for stable and efficient convergence.
- A new benchmark: MMDOD. We release the first publicly available dataset for detection-driven multi-modal and multi-view fusion, along with an adapted MVMM-DET baseline for fair comparison and reproducibility.

Extensive experiments on RGB–thermal, multi-camera, and MMDOD benchmarks show that EDIF achieves competitive performance with specialized models while uniquely unifying both fusion dimensions, demonstrating the benefits of detection-driven, object-level reasoning for robust industrial and real-world perception systems.

2 Related Work

Our work lies at the intersection of image fusion and feature matching. Image fusion aims to integrate heterogeneous sources, traditionally through either multi-modal or multi-view paradigms, while feature matching focuses on establishing explicit correspondences across inputs. EDIF unifies these two lines of research within a single detection-driven framework.

Multi-modal and multi-view fusion Multi-modal fusion integrates complementary cues from different sensing modalities such as RGB, infrared, depth, or polarization. Early approaches primarily targeted image reconstruction using encoder–decoder or generative architectures [22] [27] [26]. More recent methods adopt task-driven objectives, optimizing fusion for downstream tasks such as object detection, using attention mechanisms, uncertainty modeling, or meta-learning strategies [8] [41]. In parallel, multi-view fusion has focused on enforcing geometric consistency across cameras, initially through depth-based or 3D reasoning pipelines [33] [30], and more recently via task-driven multi-view detection frameworks such as MVDet, MVDetr, and QMVDet [13] [15]. While these approaches achieve strong performance, multi-modal methods remain largely modality-specific, and multi-view methods are typically restricted to single-modality (RGB) inputs. The joint exploitation of multi-modal and multi-view information remains comparatively underexplored. MVMM-Net [34] constitutes an early attempt to combine both dimensions, but relies on naïve feature aggregation and is not designed for detection-oriented objectives. GMMVF [6] introduced graph reasoning for multi-modal multi-view fusion, yet remains task-specific and not directly applicable to object detection. In contrast, EDIF introduces the first end-to-end detection-driven architecture that explicitly models both spectral and angular dependencies through keypoint-based graph attention matching and unified object-level reasoning.

Feature matching Feature matching methods aim to establish reliable correspondences between local features across images. SuperPoint [9] introduced self-supervised keypoint detection and description, while SuperGlue [32] incorporated graph neural networks for context-aware matching. Subsequent works such as LightGlue and OmniGlue improved efficiency and generalization [18], and recent approaches extended matching to multi-view and dense settings [10]. These advances demonstrate the effectiveness of graph-based attention for establishing consistent correspondences, motivating the matching-driven formulation adopted in EDIF.

Datasets Numerous datasets support either multi-modal fusion (e.g., KAIST, LLVIP, M3FD, MSRS) or multi-view perception (e.g., Wildtrack, MultiViewX, nuScenes, KITTI) [16] [17] [24] [37] [3] [12]. However, existing benchmarks typically isolate these two dimensions and do not provide consistent multi-modal, multi-view object-level annotations. This limitation hinders the development and evaluation of generalist multi-sensor fusion models. As emphasized in recent surveys [2], unified benchmarks remain a critical missing component, motivating the introduction of MMDOD.

3 Method

We propose EDIF, an end-to-end detection-driven framework designed to jointly address multi-view and multi-modal image fusion within a single unified architecture. Unlike classical fusion strategies that operate at the pixel or feature-map level [21, 39], EDIF formulates fusion as an object-level alignment problem, where heterogeneous observations are aggregated through keypoint correspondences and graph-based reasoning, and directly optimized for object detection following a task-driven paradigm [25, 41].

Given a set of images $\{\mathcal{I}^{(m,v)}\}$ acquired from multiple modalities $m \in \mathcal{M}$ and viewpoints $v \in \mathcal{V}$, EDIF decomposes the fusion process into three main stages: modality-agnostic keypoint extraction, graph attention-based matching and grouping of keypoints across views and modalities, and object-level decoding for bounding box regression and classification. A progressive three-stage training strategy ensures stable optimization of the full pipeline.

3.1 Keypoint Extraction with TransPoint

Each image $\mathcal{I}^{(m,v)}$ is processed by a Transformer-enhanced keypoint extractor (TransPoint), extending SuperPoint [9] with self-attention to improve robustness to modality and viewpoint variations. It outputs N keypoints:

$$\mathcal{K}^{(m,v)} = \{(\mathbf{x}_i, \mathbf{d}_i, c_i)\}_{i=1}^{N}, \tag{1}$$

where $\mathbf{x}_i$, $\mathbf{d}_i$, and c_i denote location, descriptor, and confidence. Features are extracted using a convolutional backbone followed by Transformer layers and two lightweight heads predicting heatmaps and descriptors. A top-N selection ensures a compact representation.

3.2 Graph Attention-Based Matching and Grouping

Keypoints extracted from all images are pooled into a single heterogeneous set and modeled as nodes of a fully connected graph. Each node is initialized with an embedding that concatenates descriptor and confidence information:

$$\mathbf{h}_i^{(0)} = [\mathbf{d}_i \,\|\, c_i]. \tag{2}$$

To establish cross-view and cross-modal correspondences, EDIF employs a Graph Attention Neural Network (GANN). At each layer ℓ, node embeddings are updated through multi-head attention:

$$\mathbf{h}_i^{(\ell+1)} = \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(\ell)} \mathbf{W}^{(\ell)} \mathbf{h}_j^{(\ell)}, \tag{3}$$

where $\alpha_{ij}^{(\ell)}$ are attention weights computed from pairwise similarities, and $\mathbf{W}^{(\ell)}$ is a learnable projection. This mechanism enables context-aware matching across modalities and viewpoints.

Following this refinement, keypoints are softly assigned to G latent object hypotheses using a differentiable grouping module. The assignment matrix $\mathbf{S} \in \mathbb{R}^{|\mathcal{K}| \times G}$ is obtained via Sinkhorn normalization [7], yielding an approximately permutation-invariant assignment. Each object representation is then computed as:

$$\mathbf{z}_g = \sum_i S_{i,g} \mathbf{h}_i. \tag{4}$$

This soft grouping allows each hypothesis to aggregate complementary evidence across views and modalities, producing object-centric fused representations.

3.3 Detection Head and Object-Level Decoding

Each object embedding $\mathbf{z}_g$ is decoded by a shared detection head with parallel branches for bounding box regression and classification:

$$\hat{\mathbf{b}}_g = f_{\text{box}}(\mathbf{z}_g), \quad \hat{\mathbf{p}}_g = f_{\text{cls}}(\mathbf{z}_g). \tag{5}$$

An additional objectness score filters spurious detections. This design directly predicts objects from fused representations without requiring intermediate image reconstruction.

4 Loss Function and Training Strategy

EDIF is trained under a detection-driven paradigm in which object detection provides the supervisory signal for the entire fusion pipeline, following task-driven fusion principles commonly adopted in multi-modal detection frameworks [25, 41]. Direct end-to-end optimization of all components is challenging due to the interaction between heterogeneous modules, the discrete nature of object grouping, and the variability induced by multi-view multi-modal inputs. To ensure stable convergence, we adopt a progressive three-stage training strategy, where losses of increasing semantic complexity are introduced sequentially.

4.1 Stage 1: Keypoint Localization and Descriptor Consistency

The first training stage focuses exclusively on the keypoint extractor. Since no explicit keypoint annotations are available, supervision is derived from ground-truth object locations. Each ground-truth bounding box center (x_j, y_j) is encoded as a Gaussian heatmap that acts as a weak geometric anchor:

$$T(x,y) = \sum_{j=1}^{M} \exp\left(-\frac{(x-x_j)^2 + (y-y_j)^2}{2\sigma^2}\right), \tag{6}$$

where M is the number of objects in the image and σ controls the spatial extent of the supervision.

The variance σ is progressively annealed during training, enabling a curriculum from coarse localization to fine-grained keypoint refinement, a strategy commonly used to stabilize learning under weak supervision. The keypoint detector is optimized using a combination of localization and confidence calibration losses:

$$\mathcal{L}_{\text{kp}} = \lambda_{\text{loc}} \|\hat{\mathbf{x}} - \mathbf{x}^*\|_1 + \lambda_{\text{hm}} \mathcal{L}_{\text{heatmap}}, \tag{7}$$

where $\hat{\mathbf{x}}$ denotes predicted keypoint locations and $\mathbf{x}^*$ the corresponding Gaussian targets.

To encourage view- and modality-invariant descriptors, we employ a supervised contrastive loss [20]. Keypoints associated with the same object instance across different views or modalities are treated as positives:

$$\mathcal{L}_{\text{desc}} = -\frac{1}{N} \sum_{i=1}^{N} \frac{1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp(\mathbf{f}_i^\top \mathbf{f}_j / \tau)}{\sum_{k \neq i} \exp(\mathbf{f}_i^\top \mathbf{f}_k / \tau)}, \tag{8}$$

where $\mathbf{f}_i$ denotes the descriptor of keypoint i, $P(i)$ the set of positives, and τ a temperature parameter.

This stage provides a stable initialization for downstream matching and grouping by enforcing both spatial consistency and semantic alignment at the keypoint level.

4.2 Stage 2: Object Localization and Confidence Estimation

In the second stage, the keypoint extractor, grouping module, and decoder are jointly optimized to regress object bounding boxes and objectness confidence scores. Predicted object hypotheses are matched to ground-truth objects using Hungarian assignment, with a matching cost combining ℓ_1 distance and Intersection-over-Union (IoU).

For matched predictions, bounding box regression is supervised using a mixed localization objective:

$$\mathcal{L}_{\text{box}} = \alpha \|\hat{\mathbf{B}} - \mathbf{B}^*\|_1 + (1 - \text{IoU}(\hat{\mathbf{B}}, \mathbf{B}^*)), \tag{9}$$

where $\hat{\mathbf{B}}$ and $\mathbf{B}^*$ denote predicted and ground-truth boxes, respectively.

Each object hypothesis additionally predicts an objectness score, supervised using binary cross-entropy:

$$\mathcal{L}_{\text{obj}} = -\left[y \log \hat{s} + (1-y)\log(1-\hat{s})\right], \tag{10}$$

where y indicates whether the hypothesis is matched to a ground-truth object. This loss suppresses spurious detections induced by the fixed keypoint budget and stabilizes object-level reasoning.

4.3 Stage 3: Joint Detection and Cross-Sensor Grouping

The final stage trains the full architecture end-to-end, combining detection, classification, and grouping objectives. Bounding box regression is refined using a generalized IoU formulation augmented with center and scale constraints:

$$\mathcal{L}_{\text{box}} = (1 - \text{GIoU}) + \lambda_c \|\Delta\mathbf{c}\|_1 + \lambda_s \|\Delta\mathbf{s}\|_1, \tag{11}$$

where $\Delta\mathbf{c}$ and $\Delta\mathbf{s}$ denote center and size offsets. The use of Generalized IoU improves convergence and robustness under partial overlap [42].

Object classification is supervised using Focal Loss with label smoothing to address class imbalance [23]. To explicitly enforce cross-view and cross-modal consistency, a group coherence loss is applied to the soft assignment vectors $\mathbf{S}_i$, inspired by set-based and permutation-invariant learning formulations [31]:

$$\mathcal{L}_{\text{group}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left(1 - \cos(\mathbf{S}_i, \mathbf{S}_j)\right), \tag{12}$$

where $\mathcal{P}$ denotes pairs of keypoints belonging to the same object instance.

A complementary entropy-based sparsity loss encourages peaked and interpretable assignments:

$$\mathcal{L}_{\text{sparse}} = -\frac{1}{BN} \sum_{b,n,g} S_{b,n,g} \log S_{b,n,g}. \tag{13}$$

5 MMDOD: A Multi-View and Multi-Modal Dataset for Detection

To evaluate EDIF and foster future research on image fusion, we introduce the Multi-Modal and Multi-View Object Detection Dataset (MMDOD), specifically designed for transparent object inspection. Unlike existing benchmarks that isolate either modality or viewpoint, MMDOD provides consistent multi-view, multi-modal captures with detection-oriented annotations, offering a unique resource for fusion research. MMDOD contains more than 10k images of cylindrical glass containers captured from six calibrated viewpoints under four complementary modalities: Visible, captures fine texture and colour cues, essential for

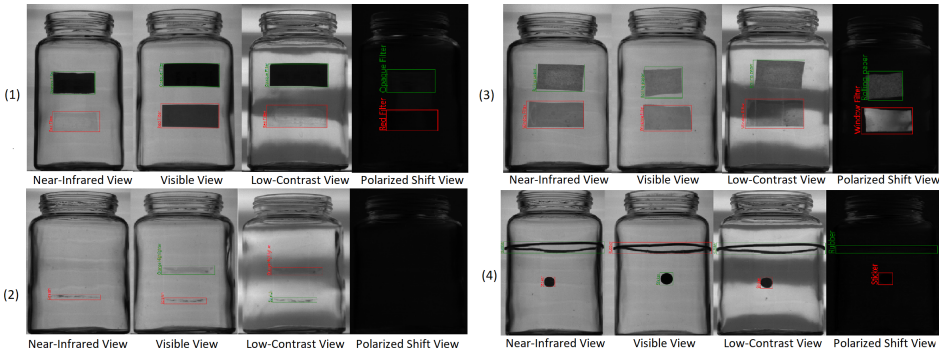

Fig. 2. Defect Appearance Under Multi-Modal Imaging. Examples of object appearances across the four imaging modalities: Near-Infrared, Visible, Low-Contrast, and Polarization Shift. (1) Red and opaque filters: visually similar in Visible and Polarized views, but clearly separable in Near-Infrared. (2) Orange highlighter and structural scratch: the highlighter shows strong modality-specific visibility, while scratches are enhanced under Low-Contrast due to refractive irregularities. (3) Rolling paper and window filter: objects of similar color are differentiated under Polarization thanks to birefringence. (4) Rubber band and colored sticker: rubber bands span large areas with variable appearance across views, whereas stickers show high optical contrast across modalities. This figure highlights how different modalities reveal complementary object characteristics, illustrating the need for joint fusion.

stickers and labels; Near-Infrared (NIR), enhances transparency defects and low-light visibility; Low-Contrast, reveals faint structural defects such as scratches or cracks; Polarization Shift, emphasizes birefringence and refractive differences, critical for transparent and reflective surfaces. The dataset includes more than 21000 annotated objects across 8 categories, covering both destructive defects (scratches) and non-destructive artifacts (stickers, rubbers, orange highlighters, opaque filters, red filters, window filters, rolling papers, elastic bands) and is split into training, validation, and test sets using a 80/20 ratio, while preserving the distribution of modalities and viewpoints. Additionally, background distractors such as mold seams, strong reflections, and shadows are present, requiring implicit negative learning. Object detectability in MMDOD is highly dependent on modality–view combinations. For instance, certain defects are invisible in RGB but clearly revealed in NIR, while birefringence effects in polarized imaging make transparent materials distinguishable. Similarly, viewpoint changes can either amplify or suppress defect visibility. This variability makes MMDOD an ideal benchmark for evaluating fusion strategies that must integrate complementary cues across both axes. Annotations are manually produced by domain experts, ensuring consistency across all views and modalities. Each object is labeled in all visible images, including partially observable instances, enabling evaluation of cross-view and cross-modal consistency. We establish baselines with MVMM-NET [34], a detection framework for multi-modal multi-view inputs, and its

stronger successor MVMM-DET, which incorporates modality-aware encoders and a more powerful fusion backbone. As reported in Table 4, EDIF consistently surpasses both, confirming the benefits of learning a unified feature space where angular and spectral cues are modelled jointly. During evaluation, predictions are matched to ground-truth objects using Hungarian assignment based on IoU and class consistency, ensuring that detections are consistent across views and modalities rather than evaluated independently. Unlike MVMM-DET, which fuses modalities and views in separate stages, EDIF performs detection-driven fusion end-to-end, leading to superior generalization. By releasing MMDOD, we provide the first benchmark explicitly tailored for generalist detection-driven fusion. This dataset establishes a standard for future research, enabling reproducibility, fair comparison, and accelerating progress toward architectures that can robustly integrate heterogeneous inputs.

6 Experimental Results

We evaluate EDIF under three complementary scenarios: multi-modal fusion (RGB–thermal), multi-view fusion (multi-camera pedestrian detection), and joint multi-modal multi-view fusion. This protocol allows us to assess whether a single detection-driven framework can remain effective across heterogeneous fusion settings typically addressed by specialized architectures. All experiments are conducted under the standard evaluation protocols of each benchmark. Experiments are implemented in PyTorch and trained using the AdamW optimizer with an initial learning rate of 10^{-4}. Training follows the proposed three-stage schedule, with a fixed number of keypoints and object hypotheses. All models are trained on NVIDIA GPUs A4000, with total training time of approximately 32 hours.

Multi-Modal Fusion We first evaluate EDIF on RGB–thermal benchmarks, which assess the ability of fusion models to exploit complementary information from visible and infrared modalities. On the M3FD dataset (Table 1), EDIF achieves an mAP@50–95 of 55.8, which is comparable to recent task-driven fusion approaches. While MetaFusion [41] reports the highest score (56.5), EDIF outperforms several widely used fusion strategies such as TarDAL [25], U2Fusion [39], and FusionGAN [27]. These results indicate that EDIF can effectively leverage spectral complementarity, despite not being specifically optimized for single multi-modal fusion.

On the FLIR dataset (Table 1), which includes challenging outdoor and low-illumination scenarios, EDIF achieves 84.2 mAP@50 and 41.2 mAP@75. While remaining slightly below the strongest specialized model (CBAM [8]), EDIF clearly outperforms MFPT-based [43] and uncertainty-aware [36] baselines. This confirms that the proposed object-level fusion strategy generalizes well to standard multi-modal detection settings.

Multi-View Fusion We next evaluate EDIF on multi-camera pedestrian detection benchmarks, which emphasize geometric consistency and cross-view reasoning.

Table 1. Object detection performance on M3FD and FLIR datasets.

M3FD (mAP@50–95)

Method (M3FD)	mAP@50–95
FusionGAN [27]	54.2
U2Fusion [39]	55.7
TarDAL [25]	54.4
MetaFusion [41]	**56.5**
EDIF (ours)	55.8

FLIR (mAP@50 & mAP@75)

Method (FLIR)	mAP@50	mAP@75
GAFF [40]	72.9	32.9
CFT [28]	78.7	35.5
CSAA [4]	79.2	37.4
LRAF-Net [11]	80.5	–
CBAM [8]	**86.2**	**43.0**
EDIF (ours)	84.2	41.2

On Wildtrack (Table 2), EDIF achieves a MODA of 0.928 and a MODP of 0.812, performing on par with recent methods such as QMVDet [15] and 3DROM [29], while outperforming earlier baselines such as MVDet [14]. EDIF also achieves one of the highest precision scores, indicating a low false-positive rate across views.

Table 2. Performance on Wildtrack.

Method	MODA	MODP	Precision	Recall
MVDet [14]	0.882	0.757	0.947	0.936
MVDetr [13]	0.915	0.821	**0.974**	0.940
3DROM [29]	**0.935**	0.759	0.972	0.962
QMVDet [15]	0.931	**0.826**	0.966	**0.964**
EDIF (ours)	0.928	0.812	0.973	0.959

On MultiViewX (Table 3), EDIF reaches a MODA of 0.947 and a MODP of 0.908, remaining close to the strongest reported methods while surpassing several established baselines. These results show that EDIF retains strong multi-view reasoning capabilities, despite not relying on architectures specifically designed for ground-plane localization.

Table 3. Performance on MultiViewX.

Method	MODA	MODP	Precision	Recall
MVDet [14]	0.893	0.796	0.968	0.867
MVDetr [13]	0.937	0.913	0.995	0.942
3DROM [29]	0.950	0.849	0.990	**0.961**
QMVDet [15]	**0.951**	**0.922**	**0.996**	0.955
EDIF (ours)	0.947	0.908	0.993	0.949

Multi-Modal and Multi-View Fusion We evaluate EDIF on the MMDOD benchmark, which requires joint exploitation of multi-modal and multi-view information. To assess the benefit of multi-modal and multi-view fusion, we compare EDIF with three baselines: a mono-image DETR [5], which processes single RGB images and represents a strong single-view detection architecture; MVMMDET [34]; and a stronger adapted variant incorporating modality-aware encoders and a more powerful backbone. Results are reported in Table 4.

Table 4. Results on MMDOD (multi-modal and multi-view).

Method	mAP@50	mAP@75	F1-Score
Baseline Mono-View DETR [5]	0.855	0.780	0.313
MVMM-DET [34]	0.649	0.292	0.321
EDIF (ours)	**0.823**	**0.655**	**0.682**

The mono-image DETR achieves high localization scores (mAP@50 and mAP@75) due to its strong single-view detection capabilities, but performs poorly on F1-Score, reflecting its difficulty in correctly classifying objects when multiple modalities and viewpoints provide complementary cues. The higher mAP of DETR can be explained by its single-view setting, which avoids cross-modal and cross-view inconsistencies. In contrast, EDIF operates in a more complex setting requiring alignment across heterogeneous inputs, which may slightly reduce localization precision as measured by mAP. In contrast, EDIF leverages the full multi-modal multi-view context, resulting in significantly improved classification performance while maintaining competitive localization. This significant gain in F1-score highlights the ability of EDIF to integrate complementary cues across sensors, improving classification robustness when objects are ambiguous or partially observable in individual views. These results demonstrate that jointly modeling spectral and angular information under detection supervision provides substantial gains over both single-view and adapted multi-modal multi-view baselines, highlighting the value of unified object-level representations.

7 Discussion and Limitations

EDIF demonstrates that a single detection-driven framework can effectively integrate multi-modal and multi-view information. On standard benchmarks, it achieves competitive performance compared to specialized methods, and on MMDOD, which explicitly requires joint spectral and angular reasoning, EDIF yields substantial improvements over adapted fusion baselines. This confirms that learning object-centric representations under unified detection supervision is particularly beneficial when both modalities and viewpoints vary. However, EDIF does not always surpass task-specific state-of-the-art models in settings limited to a single fusion dimension, reflecting the inherent trade-off between specialization and generalization. The three-stage training stabilizes optimization but increases training complexity, and the architecture incurs computational overhead from keypoint extraction and graph reasoning, limiting real-time deployment. Additionally, MMDOD focuses on a specific industrial scenario, leaving open questions about generalization to more diverse or dynamic environments. Future work includes optimizing the architecture for computational efficiency, handling missing views or modalities, and improving model components to achieve state-of-the-art performance across all fusion scenarios. Extensive studies on module contributions, hyperparameter sensitivity, and trade-offs

between accuracy and speed will further guide the design of generalist detection-driven fusion systems applicable to heterogeneous multi-sensor setups.

8 Conclusion

We presented EDIF, a end-to-end multi-view multi-modal detection-driven framework for image fusion that unifies multi-modal and multi-view settings. By leveraging graph-based feature matching, EDIF captures multi-modal and multi-view dependencies within a single representation. Experiments on RGB–thermal, multi-camera, and joint datasets show that EDIF is competitive with specialized state-of-the-art methods while uniquely generalizing across heterogeneous inputs. This flexibility positions EDIF as a promising foundation for real-world multi-sensor detection systems. In future work, we plan to extend EDIF to handle missing modalities or viewpoints, a critical challenge in real-world deployments. We aim to investigate graph neural network–based mechanisms to dynamically adapt the fusion process when certain modalities or views are unavailable. These directions hold potential to further enhance the robustness and adaptability of EDIF in unconstrained multi-sensor environments.

References

1. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence **41**(2), 423–443 (2018)
2. Bernardi, G., Brisebarre, G., Roman, S., Ardabilian, M., Dellandrea, E.: A comprehensive survey on image fusion: Which approach fits which need. Information Fusion **126**, 103594 (2026). https://doi.org/https://doi.org/10.1016/j.inffus.2025.103594, https://www.sciencedirect.com/science/article/pii/S1566253525006669
3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. arXiv preprint arXiv:1903.11027 (2019)
4. Cao, Y., Bin, J., Hamari, J., Blasch, E., Liu, Z.: Multimodal object detection by channel switching and spatial attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 403–411 (2023)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
6. Chen, J., Dey, S.: Graph-based multi-modal multi-view fusion for facial action unit recognition. IEEE Access **12**, 69310–69324 (2024)
7. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
8. 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. pp. 7366–7375 (2024)
9. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 224–236 (2018)

10. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: Roma: Robust dense feature matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19790–19800 (2024)
11. Fu, H., Wang, S., Duan, P., Xiao, C., Dian, R., Li, S., Li, Z.: Lraf-net: Long-range attention fusion network for visible–infrared object detection. IEEE Transactions on Neural Networks and Learning Systems (2023)
12. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
13. Hou, Y., Zheng, L.: Multiview detection with shadow transformer (and view-coherent data augmentation). In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1673–1682 (2021)
14. Hou, Y., Zheng, L., Gould, S.: Multiview detection with feature perspective transformation. In: European Conference on Computer Vision. pp. 1–18. Springer (2020)
15. Hsu, H.M., Yuan, X., Chuang, Y.Y., Sun, W., Chang, R.I.: Query-based multiview detection for multiple visual sensor networks. Sensors **24**(15), 4773 (2024)
16. Hwang, S., Park, J., Kim, N., Choi, Y., Kweon, I.S.: Multispectral pedestrian detection: Benchmark dataset and baselines. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
17. 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. pp. 3496–3504 (2021)
18. Jiang, H., Karpur, A., Cao, B., Huang, Q., Araujo, A.: Omniglue: Generalizable feature matching with foundation model guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19865–19875 (2024)
19. Karim, S., Tong, G., Li, J., Qadir, A., Farooq, U., Yu, Y.: Current advances and future perspectives of image fusion: A comprehensive review. Information Fusion **90**, 185–217 (2023)
20. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
21. Li, H., Wu, X.J.: Densefuse: A fusion approach to infrared and visible images. IEEE Transactions on Image Processing **28**(5), 2614–2623 (2018)
22. Li, H., Wu, X.J., Kittler, J.: Rfn-nest: An end-to-end residual fusion network for infrared and visible images. Information Fusion **73**, 72–86 (2021)
23. 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. pp. 2980–2988 (2017)
24. 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. IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 5802–5811 (2022)
25. 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. pp. 5802–5811 (2022)
26. Ma, J., Xu, H., Jiang, J., Mei, X., Zhang, X.P.: Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. IEEE Transactions on Image Processing **29**, 4980–4995 (2020)
27. Ma, J., Yu, W., Liang, P., Li, C., Jiang, J.: Fusiongan: A generative adversarial network for infrared and visible image fusion. Information fusion **48**, 11–26 (2019)

28. Qingyun, F., Dapeng, H., Zhaokui, W.: Cross-modality fusion transformer for multispectral object detection. arXiv preprint arXiv:2111.00273 (2021)
29. Qiu, R., Xu, M., Yan, Y., Smith, J.S., Yang, X.: 3d random occlusion and multi-layer projection for deep multi-camera pedestrian localization. In: European Conference on Computer Vision. pp. 695–710. Springer (2022)
30. Rabby, A., Zhang, C.: Beyondpixels: A comprehensive review of the evolution of neural radiance fields. arXiv preprint arXiv:2306.03000 (2023)
31. Rezatofighi, S.H., Kaskman, R., Motlagh, F.T., Shi, Q., Cremers, D., Leal-Taixé, L., Reid, I.: Deep perm-set net: Learn to predict sets with unknown permutation and cardinality using deep neural networks. arXiv preprint arXiv:1805.00613 (2018)
32. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4938–4947 (2020)
33. Shu, C., Deng, J., Yu, F., Liu, Y.: 3dppe: 3d point positional encoding for transformer-based multi-camera 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3580–3589 (2023)
34. Song, J., Zheng, Y., Zakir Ullah, M., Wang, J., Jiang, Y., Xu, C., Zou, Z., Ding, G.: Multiview multimodal network for breast cancer diagnosis in contrast-enhanced spectral mammography images. International Journal of Computer Assisted Radiology and Surgery **16**(6), 979–988 (2021)
35. Sun, J., Chen, L., Xie, Y., Zhang, S., Jiang, Q., Zhou, X., Bao, H.: Disp r-cnn: Stereo 3d object detection via shape prior guided instance disparity estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10548–10557 (2020)
36. Sun, Y., Cao, B., Zhu, P., Hu, Q.: Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. IEEE Transactions on Circuits and Systems for Video Technology **32**(10), 6700–6713 (2022)
37. Tang, L., Yuan, J., Zhang, H., Jiang, X., Ma, J.: Piafusion: A progressive infrared and visible image fusion network based on illumination aware. Information Fusion (2022)
38. Vs, V., Valanarasu, J.M.J., Oza, P., Patel, V.M.: Image fusion transformer. 2022 IEEE International Conference on Image Processing (ICIP) pp. 3566–3570 (2022)
39. Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H.: U2fusion: A unified unsupervised image fusion network. IEEE transactions on pattern analysis and machine intelligence **44**(1), 502–518 (2020)
40. Zhang, H., Fromont, E., Lefèvre, S., Avignon, B.: Guided attentive feature fusion for multispectral pedestrian detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 72–80 (2021)
41. Zhao, W., Xie, S., Zhao, F., He, Y., Lu, H.: Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13955–13965 (2023)
42. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-iou loss: Faster and better learning for bounding box regression. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12993–13000 (2020)
43. Zhu, Y., Sun, X., Wang, M., Huang, H.: Multi-modal feature pyramid transformer for rgb-infrared object detection. IEEE Transactions on Intelligent Transportation Systems **24**(9), 9984–9995 (2023)