Self-Supervised Alignment of RGB-Infrared Representations for Embedded Perception

Abdelmalek Belghomari[1,2], Clara Barbanson[1], Frederic Jurie[2], Alexis Lechervy[2]

[1] Safran Electronics and Defense
[2] Université de Caen Normandie, GREYC

8 mars 2026

Résumé

La fusion d'images RVB-IR exige un alignement précis et des bases de données annotées. Pour s'en affranchir, nous proposons une approche auto-supervisée basée sur l'architecture JEPA. En prédisant les caractéristiques latentes de l'IR à partir d'un contexte RVB masqué, notre modèle projette les deux modalités dans un espace sémantique partagé. Les résultats préliminaires montrent que cet alignement constitue une base solide pour la perception embarquée, sans aucune intervention humaine.

Mots-clés

Apprentissage auto-supervisé, JEPA, Alignement de domaines, Multimodalité, Perception embarquée.

Abstract

RGB and IR image fusion requires precise alignment and annotated datasets. To eliminate this need for manual labeling, we propose a self-supervised approach using the Joint-Embedding Predictive Architecture (JEPA). By predicting IR latent features from masked RGB context, our model projects both modalities into a shared semantic space. Preliminary results show this alignment provides a solid foundation for embedded perception without any human intervention.

Keywords

Self-Supervised Learning, JEPA, Domain Alignment, Multimodality, Embedded Perception.

1 Introduction

Aligning visible (RGB) and infrared (IR) domains is a prerequisite for robust multispectral perception in defense and embedded systems. While RGB sensors excel in providing high-resolution textures and contextual details during the day, IR sensors capture thermal signatures that remain highly reliable in degraded visual environments, such as nighttime, dense fog, or camouflage scenarios. However, designing an effective fusion strategy that respects the strict power and latency constraints of embedded hardware remains a significant challenge.

Existing fusion strategies are typically categorized by their integration stage and architectural backbone. Early and middle fusion methods concatenate inputs or intermediate representations, but require extensive paired data for precise alignment. Late fusion processes modalities independently ; while effective on popular benchmarks, the dual-backbone overhead severely limits embedded deployment. Regarding backbones, Transformers (e.g., CAFF-DINO [4]) utilize cross-attention for state-of-the-art accuracy, though their quadratic complexity hinders real-time use. Alternatively, State-Space Models (SSMs) like Mamba [2] and its vision variants [6] offer linear complexity, but their efficacy for multimodal fusion remains an open empirical question.

Consequently, several key limitations emerge from these dominant approaches in the embedded context : *(1) Annotation dependency :* Most early and middle fusion methods require dense labeled RGB-IR pairs, which are costly and difficult to acquire in operational defense conditions. *(2) Computational cost :* Cross-attention and dual-backbone architectures are computationally expensive, making them challenging to run at real-time frame rates on embedded SoCs or FPGAs. *(3) Scene generalization :* Supervised models trained on specific benchmarks like LLVIP [5] often fail to generalize to new operational environments without extensive retraining.

To address these challenges, we argue that self-supervised learning (SSL) offers a practical path toward efficient cross-modal alignment. In this paper, we propose a research direction based on the Joint-Embedding Predictive Architecture (JEPA) [1] as a zero-annotation framework to learn a shared latent space, providing a lightweight and scene-agnostic foundation for real-time embedded perception.

2 Proposed Methodology

The primary motivation for adopting an SSL framework lies in the critical scarcity of annotated multimodal data. By leveraging unlabelled raw sensor data, we shift the bottleneck from expert human annotation to the simple acquisition of raw data. To this end, we propose a self-supervised pretraining strategy based on JEPA [1], adapted specifically for cross-modal alignment.

2.1 Cross-Modal JEPA Formulation

JEPA learns representations by predicting the latent representation of a target view from a context view, *entirely in*

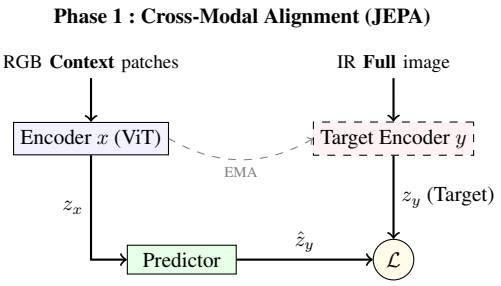

FIGURE 1 – Cross-modal Latent Alignment via JEPA

feature space. This is a critical distinction from masked autoencoders (MAE) [3] : predicting in latent space forces the model to capture high-level semantic structures rather than low-level pixel statistics, which is ultimately better suited for downstream detection tasks.

We adapt JEPA to the cross-modal RGB-IR setting through three main components : *The Context Encoder* processes patches of the RGB image. Crucially, a large portion of these RGB tokens are masked. The encoder's objective is to learn a compact visible-spectrum representation from the remaining visible context. *The Target Encoder* receives the corresponding, strictly aligned IR image. The weights of the target encoder are an exponential moving average (EMA) of the context encoder. This prevents representation collapse and provides stable cross-modal latent targets. *The Predictor* acts as the alignment mechanism. Taking the encoded RGB context and the positional embeddings of the masked tokens, it is trained to predict the exact latent features of the corresponding masked regions in the IR domain. The entire framework is optimized by minimizing the L_2 distance between the predictor's output and the target encoder's representation.

2.2 Methodological Advantages for Embedded Systems

Our proposed formulation directly addresses embedded limitations : *(1) Zero-annotation fusion :* It relies solely on co-registered RGB-IR pairs, completely eliminating the need for bounding boxes or semantic maps during pretraining. *(2) Lightweight design :* By using a lightweight Vision Transformer (ViT-Tiny) with approximately 5.7M parameters, the backbone remains strictly compatible with embedded hardware targets like low-power edge devices. *(3) Domain Agnosticism :* The SSL objective relies on structural cross-modal correlation rather than class labels, theoretically allowing the network to be trained on diverse unlabelled operational datasets (desert, jungle, urban) without manual retagging.

3 Preliminary Proof of Concept & Future Work

To validate our proposed methodology, we trained the cross-modal JEPA on the LLVIP dataset [5], which contains over 15,000 aligned image pairs recorded in challenging low-light conditions. We utilized a ViT-Tiny backbone with a patch size of 16 and a resolution of 224×224, optimizing the architecture over 450 epochs to ensure full convergence of the self-supervised representations.

To evaluate the learned representation, we froze the pretrained encoder and attached a DETR (DEtection TRansformer) head. Because DETR relies heavily on rich, global feature sets for its bipartite matching algorithm, it serves as an excellent probe for feature quality. At epoch 450, our model reached a stable localization plateau ($L_1 \approx 0.051$, $IoU \approx 67.5\%$), proving that the cross-modal self-supervised features capture the spatial geometry necessary for object detection. However, a classification error of 60% indicates that capturing fine-grained semantic distinctions remains an ongoing challenge.

Open Research Directions : Building on these preliminary insights, several methodological paths remain open : *(1) Information Saturation :* The model likely saturates the semantic information available in static LLVIP pairs. We plan to implement aggressive cross-modal data augmentations to synthetically expand the training distribution. *(2) SSM Alternatives :* Investigating whether State-Space Models (e.g., VMamba) can replace ViT within the JEPA framework to further reduce inference complexity for extreme embedded targets. *(3) Transferability :* Evaluating whether this scene-agnostic pretraining inherently transfers better to unobserved operational environments compared to standard supervised baselines.

4 Conclusion

We proposed a self-supervised methodology using JEPA to establish a shared RGB-IR semantic space without manual labels. By predicting IR latent features from masked RGB contexts, the framework intrinsically aligns both modalities. Preliminary validation confirms stable geometric feature extraction using a lightweight ViT-Tiny backbone. Future work will refine semantic classification via cross-modal augmentations and explore sub-quadratic SSM backbones to satisfy extreme embedded perception constraints.

Références

[1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023.

[2] Albert Gu and Tri Dao. Mamba : Linear-time sequence modeling with selective state spaces. In *ICLR*, 2024.

[3] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[4] Kevin Helvig, Baptiste Abeloos, and Pauline Trouvé-Peloux. CAFF-DINO : Multi-spectral object detection transformers with cross-attention features fusion. In *CVPR*, 2024.

[5] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. LLVIP : A visible-infrared paired dataset for low-light vision. In *ICCV Workshops*, 2021.

[6] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunnan Liu. VMamba : Visual state space model. In *NeurIPS*, 2024.