DenseDeCUR: Dense, Decoupling of Common and Unique Representations

Daniele Kopyshevskiy Davide Secco Marco Casagrande
Politecnico di Milano
{daniele.kopyshevskiy, davide.secco, marco.casagrande}@mail.polimi.it

Abstract

Multimodal data has become increasingly important across various computer vision tasks, as combining information from different modalities often leads to superior performance. Self-supervised learning has emerged as a promising paradigm in this context, enabling the exploitation of large-scale unlabeled multimodal data to learn rich representations without manual annotation. This work presents DenseDeCUR, a multimodal self-supervised learning method that learns dense feature maps while decoupling shared (common) and modality-specific (unique) representations at the patch level. We achieve this by combining DeCUR's multimodal redundancy reduction framework with DenseCL's dense contrastive representation learning, enabling the model to distinguish between inter- and intra-modal embeddings at a fine-grained spatial resolution. Compared to the baseline methods, our approach demonstrates superior performance when transferred to downstream dense prediction tasks. Specifically, on the KAIST multispectral pedestrian detection dataset, DenseDeCUR achieves 50.06% mAP@0.5 and 17.97% mAP@0.5:0.95, outperforming DeCUR by 1.55 and 0.86 percentage points respectively. These results demonstrate the strong potential of combining dense representations with explicit modality decoupling for multimodal self-supervised learning. Temporary repository available at: `https://github.com/DavideSecco/ADL-Project`

1. Introduction

Humans perceive the world through multiple senses, including sight, sound, and touch, obtaining a comprehensive understanding by leveraging complementary information from all these modalities [13]. Similarly, in machine learning, multimodal approaches have demonstrated significant advantages over single-modal methods across various applications such as autonomous driving, video understanding, and visual question answering. In self-supervised learning, two major research lines exist: single-modal and multimodal learning [6]. Single-modal methods typically learn global image-level representations suited for classification but lacking spatial structure for dense prediction tasks. DenseCL [9] addressed this by extending contrastive learning to pixel-level, learning dense feature maps. Multimodal methods aim to align representations from different modalities. While most learn only common representations, DeCUR [10] explicitly decouples common and unique components of each modality. However, it operates on global vectors rather than dense features. We propose DenseDeCUR, a multimodal self-supervised framework that *i)* learns intra-modal agreements between feature vectors via dense contrastive learning, *ii)* factorizes each feature vector into shared and unique components, and *iii)* enforces cross-modal alignment on shared components while decorrelating unique components to preserve modality-specific information. In summary, our main contributions are:

- We propose DenseDeCUR, performing dense patch-level contrastive learning in a multimodal framework with decoupled representations.
- We demonstrate that patch-level learning within a multimodal decoupling framework yields better results than global representations for dense prediction tasks.

2. Related Work

In visual recognition, the traditional pre-training paradigm has been dominated by supervised learning on ImageNet [2] mainly for image classification tasks. Initial self-supervised approaches adopted a similar focus. Among them, SimCLR [1] introduced a contrastive learning framework that maximizes similarity between positive pairs while minimizing similarity to negative samples. MoCo [4] addressed computational limitations by introducing a momentum-updated encoder and queue-based dictionary, enabling larger numbers of negative samples without large batch sizes. MoCo v2 [3] adopted key components from SimCLR including an MLP projection head and stronger augmentation. However, these methods yielded

limited results on dense downstream tasks compared to supervised learning. DenseCL [9] addressed this by reformulating contrastive learning at the pixel level, operating on dense feature maps instead of global vectors. In parallel, redundancy-reduction methods offered an alternative to contrastive learning. Barlow Twins [12] introduced a cross-correlation objective that pushes the correlation matrix between embeddings toward identity: diagonal entries toward one (correlated features) and off-diagonal entries toward zero (decorrelation), reducing redundancy. DeCUR [10] extended this principle to multimodal learning. At the unimodal level, it applies a Barlow Twins–style loss within each modality. At the cross-modal level, embeddings are split into common and unique subspaces: the correlation matrix is partitioned, with the common block pushed toward identity (alignment) and the unique block toward zero (decorrelation).

3. Problem Formulation

We consider a self-supervised learning setting for multimodal data. The input is a dataset $\mathcal{D} = \{(x_i^{m_1}, x_i^{m_2})\}_{i=1}^N$ consisting of unlabeled image pairs from two distinct modalities $m \in \{m_1, m_2\}$. The output consists of two learned modality-specific encoders f_{m_1} and f_{m_2}. Each encoder $f_m : x_i^m \mapsto F_i^m$ maps an input image to a dense feature map. The goal is to learn dense representations for both modalities that capture semantically meaningful information suitable for downstream dense prediction tasks, such as object detection and semantic segmentation.

4. Methods

Figure 1 presents the general architecture of DenseDeCUR. At a high level, our framework consists of two parallel branches, one for each modality. Inspired by DeCUR, our training objective is decomposed into two components: intra-modal and cross-modal learning. The complete training objective combines these components as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{intra}}^{m_1} + \mathcal{L}_{\text{intra}}^{m_2} + \mathcal{L}_{\text{cross}}, \tag{1}$$

where $\mathcal{L}_{\text{intra}}^{m_1}$ and $\mathcal{L}_{\text{intra}}^{m_2}$ are the intra-modal losses for modalities m_1 and m_2 respectively (defined in Sec. 4.1), $\mathcal{L}_{\text{cross}}$ is the cross-modal loss (defined in Sec. 4.2). To achieve dense representations with explicit modality decoupling, we modify DeCUR [10] to operate on dense feature maps rather than (only) on global feature vectors by incorporating the principles of DenseCL [9]. More in detail, for each image x^m of modality m, two random views $(x^m)'$ and $(x^m)''$ are generated by random data augmentation. For simplicity, we omit the superscript m in the following notation. Following MoCo [4], each pair of augmented views is processed by two encoders: a query encoder f^q and a key encoder f^k. View x' is processed by f^q while view x'' is processed by f^k. Therefore, in total we have four encoders: two query encoders (one per modality) updated via backpropagation, and two key encoders (one per modality) updated via momentum. Each encoder is followed by a projection head g consisting of two parallel sub-heads: a global head g^g that produce global feature representations, and a dense head g^d that generates dense feature maps. The global head enables learning holistic image-level features, while the dense head preserves spatial structure necessary for downstream dense prediction tasks.

4.1. Intra-Modal Learning

Intra-modal learning is applied independently to the two augmented views of the same modality, and this process is performed for both modalities m_1 and m_2. The purpose of intra-modal learning is to enforce consistency between different views of the same image, learning invariant representations that capture semantic content regardless of the specific augmentation applied. To achieve this, we adopt the DenseCL framework [9], which performs contrastive learning at both the global and dense levels through its dense contrastive loss. This approach utilizes both global and dense projection heads, explained in detail below.

Global Projection Head. The global heads takes feature maps from the encoder and generates global feature vectors. Let $q = g^g(f^q(x'))$ be the encoded query and $k = g^g(f^k(x''))$ be the encoded key, with $q, k \in \mathbb{R}^K$. For each query q, there is a set of keys $\mathcal{K} = \{k_0, k_1, \dots\}$ s.t. $|\mathcal{K}| = N$, with one positive key $k_+ = k$ (global feature vector of a different view of the same image) and negative keys $\{k_-\}$ (global feature vectors of different views of different images) s.t. $\{k_-\} = \mathcal{K} \backslash \{k_+\}$. The global contrastive loss is the InfoNCE loss defined as:

$$\mathcal{L}_g = -\log \frac{\exp(q \cdot k_+/\tau)}{\exp(q \cdot k_+) + \sum_{k_-} \exp(q \cdot k_-/\tau)}, \tag{2}$$

where τ is the temperature parameter.

Dense Projection Head. The dense head takes feature maps from the encoder and generates dense feature maps $\Theta^q = g^d(f^q(x'))$ and $\Theta^k = g^d(f^k(x''))$, with $\Theta^q, \Theta^k \in \mathbb{R}^{S \times S \times K}$. For each spatial location $s \in \{1, \dots, S^2\}$, let $r^s \in \mathbb{R}^K$ denote the encoded query extracted from Θ^q at location s. Due to independent augmentations, the corresponding semantic region in Θ^k may appear at a different spatial location c_s. Suppose we have already identified the corresponding location c_s for each query location s. Let $t^{c_s} \in \mathbb{R}^K$ denote the encoded key extracted from Θ^k at location c_s. For each query r^s, there is a set of keys $\mathcal{T}^s = \{t_0^s, t_1^s, \dots\}$ s.t. $|\mathcal{T}^s| = N$, with one positive key

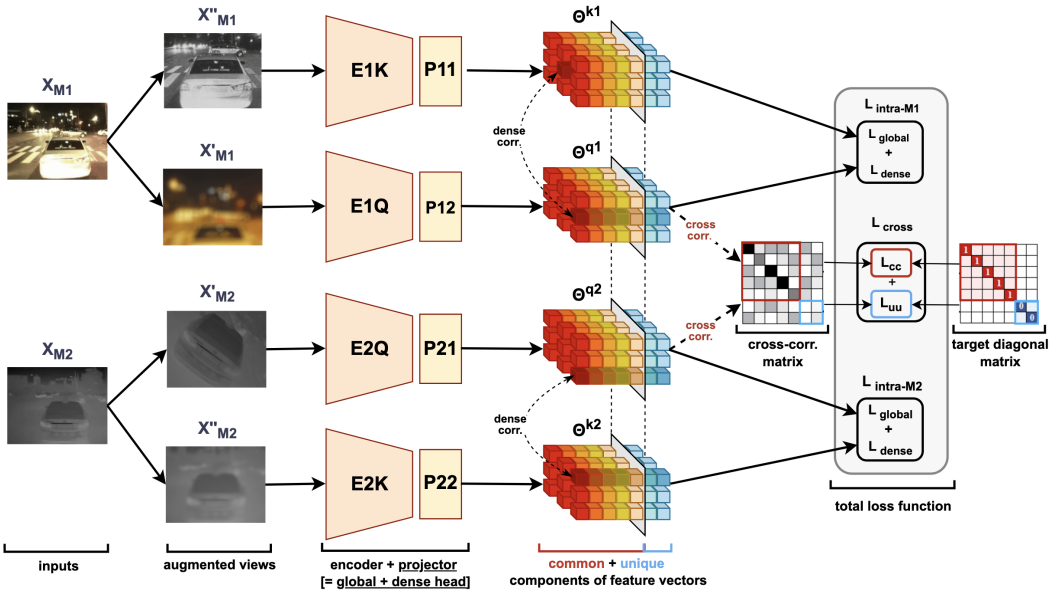

Figure 1. Architecture of DenseDeCUR. For each modality $m \in \{m_1, m_2\}$, two augmented views are processed by a query encoder (E1Q, E2Q) and a key encoder (E1K, E2K), followed by projection heads with parallel sub-heads (P11, P12, P21, P22). Each projection head consists of a global head (producing global representations) and a dense head (producing dense feature maps Θ^k and Θ^q). For visualization clarity, only the outputs of the dense heads are shown. The feature vectors in each dense feature map are decomposed into common (red) and unique (blue) components. Intra-modal learning establishes dense correspondence between Θ^k and Θ^q within each modality and applies contrastive losses ($\mathcal{L}_{\text{global}} + \mathcal{L}_{\text{dense}}$) to both global and dense representations. Cross-modal learning computes a cross-correlation matrix between spatially corresponding feature vectors from the two modalities. The common block (C^{cc}) is optimized toward identity (alignment), while the unique block (C^{uu}) is pushed toward zero (decorrelation). The target diagonal matrix visualizes the desired correlation structure. The total loss function combines intra-modal losses for both modalities and the cross-modal loss.

$t^s_+ = t^{c_s}$ (feature vector corresponding to the same semantic region in a different view of the same image) and negative keys $\{t^s_-\} = \mathcal{T}^s \setminus \{t^s_+\}$ (globally pooled vectors from dense feature maps of different view of different images). The dense contrastive loss is:

$$\mathcal{L}_d = \frac{1}{S^2} \sum_{s=1}^{S^2} -\log \frac{\exp(r^s \cdot t^s_+ / \tau)}{\exp(r^s \cdot t^s_+) + \sum_{t^s_-} \exp(r^s \cdot t^s_- / \tau)}, \tag{3}$$

where τ is the temperature parameter.

Intra-Modal Loss. The final intra-modal loss combines both global (2) and dense contrastive losses (3) as a weighted sum:

$$\mathcal{L}_{\text{intra}} = (1 - \lambda)\mathcal{L}_g + \lambda \mathcal{L}_d, \tag{4}$$

where $\lambda \in [0, 1]$ balances the contribution of $\mathcal{L}_g$ and $\mathcal{L}_d$. This loss is computed for each modality m_1 and m_2 independently. Therefore, there are two intra-modal losses, respectively $\mathcal{L}^{m_1}_{\text{intra}}$ and $\mathcal{L}^{m_2}_{\text{intra}}$.

Dense Correspondence. As mentioned above, due to independent augmentations applied to x' and x'', semantically correspondent regions may appear at different spatial locations in Θ^q and Θ^k. We denote with $\{\vartheta^q_1, \ldots, \vartheta^q_{S^2}\}$ and $\{\vartheta^k_1, \ldots, \vartheta^k_{S^2}\}$ the sequences of S^2 feature vectors of respectively Θ^q and Θ^k, where $\vartheta^q_i, \vartheta^k_j \in \mathbb{R}^K$, with $i, j = 1, \ldots, S^2$. Then, we compute a pairwise similarity matrix $\Delta \in \mathbb{R}^{S^2 \times S^2}$, where each entry Δ_{ij}, defined as

$$\Delta_{ij} = \frac{(\vartheta^q_i)^\top \vartheta^k_j}{(||\vartheta^q_i|| \cdot ||\vartheta^k_j||)}, \tag{5}$$

represents the (cosine) similarity between the feature vectors at spatial locations i and j. For each spatial location i in Θ^q, we find its corresponding spatial location in Θ^k by c_i s.t.

$$c_i = \arg \max_{j \in \{1, \ldots, S^2\}} \Delta_{ij}. \tag{6}$$

This correspondence defines intra-modal dense positive pairs: the encoded query r^i at location i in Θ^q is paired (semantically) with t^{c_i} at location c_i in Θ^k, i.e., $t^i_+ = t^{c_i}$.

4.2. Cross-Modal Learning

Cross-modal learning aims to establish semantic correspondences between the two modalities and align their

shared information while preserving modality-specific characteristics. Unlike intra-modal learning which operates within a single modality, cross-modal learning operates across modalities to exploit their complementary nature.

Modality Decoupling. For cross-modal learning, we consider the dense feature maps produced by the key encoders Θ^{k,m_1} and Θ^{k,m_2} (since they are updated with backpropagation). For simplicity, we call them Θ^{m_1} and Θ^{m_2} respectively. For each feature vector, we decompose the embedding dimensions into two disjoint subsets as in DeCUR [10]: a common (shared) subspace of dimension K_c representing information shared across modalities, and a unique subspace of dimension K_u capturing modality-specific information, where $K = K_c + K_u$. Therefore, the feature vector at spatial location i of Θ^{m_1} and at spatial location j of Θ^{m_2} can be partitioned as:

$$\vartheta_i^{m_1} = \begin{bmatrix} c_i^{m_1} \\ u_i^{m_1} \end{bmatrix} \in \mathbb{R}^K, \quad \vartheta_j^{m_2} = \begin{bmatrix} c_j^{m_2} \\ u_j^{m_2} \end{bmatrix} \in \mathbb{R}^K, \quad (7)$$

where $c_i^{m_1}, c_j^{m_2} \in \mathbb{R}^{K_c}$ and $u_i^{m_1}, u_j^{m_2} \in \mathbb{R}^{K_u}$.

Cross-Modal Correspondence. To establish correspondences between spatial locations across modalities, we use only the common component. We denote with $\{c_1^{m_1}, \ldots, c_{S^2}^{m_1}\}$ and $\{c_1^{m_2}, \ldots, c_{S^2}^{m_2}\}$ the sequences of S^2 common feature vectors extracted from respectively Θ^{m_1} and Θ^{m_2}. Following the same procedure described as above, we compute a pairwise similarity matrix $\Delta^{\text{cross}} \in \mathbb{R}^{S^2 \times S^2}$ based on the cosine similarity between common feature vectors. For each spatial location i in Θ^{m_1}, we find its corresponding location j_i^* in Θ^{m_2} by selecting the location with maximum similarity. This correspondence defines cross-modal dense positive pairs: the feature vector $\vartheta_i^{m_1}$ at location i in Θ^{m_1} is paired (semantically) with $\vartheta_{j_i^*}^{m_2}$ at location j_i^* in Θ^{m_2}. The common component is used for correspondence because it represents the only subspace where the two modalities share the same semantic space. The unique component contains modality-specific information that should not be aligned across modalities.

Cross-Correlation Matrix. Once cross-modal dense positive pairs are established, we compute the cross-correlation C between the complete feature vectors (both common and unique components) from the two modalities. Following the Barlow Twins [12] formulation, for each pair of dimensions $p, q \in \{1, \ldots, K\}$, we aggregate over all corresponding spatial location pairs across the batch:

$$C_{pq} = \frac{\sum_{b=1}^{B} \sum_{i=1}^{S^2} \vartheta_{b,i,p}^{m_1} \cdot \vartheta_{b,j_i^*,q}^{m_2}}{\sqrt{\sum_{b=1}^{B} \sum_{i=1}^{S^2} (\vartheta_{b,i,p}^{m_1})^2} \sqrt{\sum_{b=1}^{B} \sum_{i=1}^{S^2} (\vartheta_{b,j_i^*,q}^{m_2})^2}}, \quad (8)$$

where B is the batch size, j_i^* denotes the corresponding spatial location in modality m_2 for location i in modality m_1 as established by the cross-modal correspondence procedure described above, and $\vartheta_{b,i,p}^{m}$ extracts the p-th dimension of the feature vector at spatial location i in the b-th image of modality m. The resulting cross-correlation matrix $C \in \mathbb{R}^{K \times K}$ can be partitioned into blocks according to the embedding decomposition:

$$C = \begin{bmatrix} C^{cc} & C^{cu} \\ C^{uc} & C^{uu} \end{bmatrix}, \quad (9)$$

where $C_{cc} \in \mathbb{R}^{K_c \times K_c}$ captures correlations between common components, $C_{uu} \in \mathbb{R}^{K_u \times K_u}$ captures correlations between unique components, and C_{cu}, C_{uc} represent cross-block correlations.

Cross-Modal Loss. Following DeCUR [10], we apply different objectives to different blocks of the cross-correlation matrix. On one hand, to learn cross-modal common information the Barlow Twins loss is applied on C^{cc}:

$$\mathcal{L}_{cc} = \sum_{i=1}^{K_c} (1 - C_{ii}^{cc})^2 + \lambda_c \cdot \sum_{i=1}^{K_c} \sum_{j \neq i}^{K_c} (C_{ij}^{cc})^2. \quad (10)$$

The first term is the invariance term (makes each common feature dimension invariant to the input modalities) and the second term is the redundancy reduction term (decorrelates different feature dimensions in the common subspace), whereas λ_c is a positive trade-off constant. One the other hand, to decouple modality-specific information following loss is applied:

$$\mathcal{L}_{uu} = \sum_{i=1}^{K_u} (C_{ii}^{uu})^2 + \lambda_u \cdot \sum_{i=1}^{K_u} \sum_{j \neq i}^{K_u} (C_{ij}^{uu})^2. \quad (11)$$

The first term is the decorrelation term (decorrelates different modalities) and the second term is the redundancy reduction term (decorrelates different feature dimensions in the unique subspace), whereas λ_u is a positive trade-off constant. Therefore, the complete cross-modal loss combines both terms 10 and 11:

$$\mathcal{L}_{\text{cross}} = \mathcal{L}_{cc} + \mathcal{L}_{uu}. \quad (12)$$

5. Implementation Details

Dataset. All experiments were conducted on the KAIST Multispectral Pedestrian Detection dataset [7], which comprises 95,324 aligned RGB-thermal image pairs at 640×480 resolution, with 103,128 annotated pedestrian instances across person, people, and cyclist classes. The dataset spans 11 sets (set00–set11) organized into video sequences, capturing diverse illumination conditions (50.7% day, 49.3%

night). Following the official split, sets 06–11 (45,140 pairs) were reserved for testing. The official training split (sets 00–05, 50,184 pairs) was further partitioned by alternately assigning complete video sequences regardless of their set membership or sequence number: alternating videos formed the pretraining subset (23,349 pairs) and the detection fine-tuning subset (26,835 pairs), respectively. This video-level partitioning preserved temporal coherence while enabling independent evaluation of self-supervised representation learning and downstream transfer performance.

Image Augmentations. For each modality, two augmented views were generated independently using modality-specific transformation pipelines. RGB images underwent random resized cropping, random brightness and contrast adjustment, random grayscale conversion, Gaussian blur, and random horizontal flipping, followed by ImageNet normalization. Thermal images received a simplified augmentation scheme consisting of random resized cropping with identical scale parameters, random horizontal flipping, and tensor normalization without mean-std standardization to preserve the distinct statistical properties of thermal imagery.

Model Architecture. Our framework employs two independent branches, one per modality. Each encoder is ResNet-50 [5] pretrained on ImageNet [2], followed by a projection head with dual output pathways. The encoder outputs 2048-dimensional feature maps at 7×7 spatial resolution. The projection head comprises: (i) a global pathway with adaptive average pooling and a two-layer MLP (2048-2048-128) producing image-level representations, and (ii) a dense pathway with two convolutional layers (2048-2048-128, 1×1 kernel-size) preserving spatial structure and generating 128×49 dense feature maps. Following MoCo [4], we maintain momentum-updated key encoders ($m = 0.999$) and dictionary queues of size $N = 65{,}536$ for both global and dense representations. The 128-dimensional embedding space is partitioned into common ($K_c = 96$) and unique ($K_u = 32$) subspaces. For contrastive learning, we set temperature $\tau = 0.1$, while the dense-global loss weighting is $\lambda = 0.5$. The cross-modal Barlow Twins loss [12] uses redundancy reduction coefficients $\lambda_c = \lambda_u = 0.0051$ for both common and unique blocks.

6. Experiments

6.1. DenseDeCUR Pretrain

We pretrain DenseDeCUR for 200 epochs with a batch size of 128 on 1 NVIDIA A100 GPU. Images are resized to 224×224 pixels during augmentation. For optimization, we employ LARS [11] with momentum 0.9, weight decay 10^{-4}, and trust coefficient $\eta = 0.001$. The base learning rates are set to $\alpha_w = 0.002$ for weights and $\alpha_b = 0.00048$ for biases. We apply a 10-epoch linear warmup followed by cosine annealing decay over the remaining epochs. Weight decay and LARS adaptation are selectively applied, excluding bias and normalization parameters to prevent over-regularization. We also evaluate a variant using SGD optimizer with initial learning rate 0.015, momentum 0.9, and weight decay 10^{-4}, applying the same warmup and decay schedule, as in Table 1.

6.2. Fine-tuning model (Object Detection)

In our study, ICAFusion [8] was adopted as the downstream detection framework to validate the effectiveness of our self-supervised pretraining approach. ICAFusion is a feature fusion module based on dual cross-attention transformers, designed specifically for multispectral object detection. It models global interactions between modalities (such as RGB and thermal images) and aggregates complementary information across channels, using an iterative mechanism to efficiently share parameters and reduce computational complexity. The training of ICAFusion was conducted starting from different checkpoints, each corresponding to a distinct self-supervised pretraining process. This allowed us to evaluate how varying the pretrained representations impacted downstream detection performance. We trained ICAFusion for 12 epochs using 4 NVIDIA A100 GPUs, adhering to the standard training hyperparameters provided by the original ICAFusion implementation. This setup ensured a fair and consistent comparison of the results derived from different self-supervised pretrained initializations.

6.3. Results

Table 1 presents the object detection performance on the KAIST test set after fine-tuning ICAFusion with different pretrained encoders.

Baseline Comparisons. The "None" baseline refers to ICAFusion trained with ImageNet-pretrained ResNet-50 encoders without any additional self-supervised pretraining, achieving 40.88% mAP@0.5 and 14.11% mAP@0.5:0.95. DeCUR, which operates on global representations with cross-modal decoupling, obtains 48.51% mAP@0.5 and 17.11% mAP@0.5:0.95, demonstrating the effectiveness of explicit modality decoupling. DenseCL, which learns dense representations through contrastive learning but without cross-modal alignment, was pretrained separately on RGB and thermal modalities. The resulting model achieves 45.29% mAP@0.5 and 16.25% mAP@0.5:0.95, showing that dense representations alone provide limited benefits without proper multimodal integration.

Table 1. Object detection performance on KAIST test set (45,140 images, 56,336 labels). TP: True Positives, FP: False Positives, FN: False Negatives, P: Precision, R: Recall.

MODELS	TP	FP	FN	F1	P	R	MAP@.5	MAP@.5:.95
NONE	24990	**17540**	31350	0.5055	0.5876	0.4436	0.4088	0.1411
DECUR	29270	18770	27070	0.5608	0.6093	0.5195	0.4851	0.1711
DENSECL	29300	19560	27030	0.5571	0.5997	0.5201	0.4529	0.1625
DENSEDECUR-SGD (BASE)	29700	18890	26640	0.5661	0.6113	0.5272	0.4759	0.1726
DENSEDECUR-SGD	**30800**	21240	**25530**	0.5684	0.5918	**0.5468**	0.4872	0.1769
DENSEDECUR-LARS (BASE)	30520	19700	25820	0.5728	0.6077	0.5417	0.4723	0.1741
DENSEDECUR-LARS	30690	18850	25650	**0.5797**	**0.6195**	0.5447	**0.5006**	**0.1797**

DenseDeCUR Variants. We evaluate both baseline and full versions of DenseDeCUR with two optimizers. The baseline version ("base") applies Barlow Twins loss to globally pooled vectors from dense feature maps, similar to DeCUR's approach but starting from dense representations. In contrast, the full version applies the cross-modal loss directly to spatially corresponding dense feature vectors, enabling explicit patch-level modality decoupling. DenseDeCUR-SGD (base) achieves 47.59% mAP@0.5 and 17.26% mAP@0.5:0.95, while the full version reaches 48.72% mAP@0.5 and 17.69% mAP@0.5:0.95. The full version shows improved recall (54.68% vs 52.72%) with a slight increase in false positives, indicating better detection of challenging instances. DenseDeCUR-LARS consistently outperforms the SGD variant. The baseline version obtains 47.23% mAP@0.5 and 17.41% mAP@0.5:0.95, while the full version achieves the best overall results: 50.06% mAP@0.5 and 17.97% mAP@0.5:0.95. Compared to DeCUR, DenseDeCUR-LARS improves mAP@0.5 by 1.55 percentage points and mAP@0.5:0.95 by 0.86 percentage points. Notably, it also maintains the lowest false positive rate (18,850) among high-performing models while achieving the second-highest true positives (30,690) and a superior precision-recall balance (F1=0.5797, P=0.6195).

7. Conclusions

This work presented DenseDeCUR, a multimodal self-supervised learning framework that extends modality decoupling to dense feature maps through patch-level contrastive learning. Experimental results on the KAIST multispectral pedestrian detection dataset demonstrate that our approach achieves 50.06% mAP@0.5 and 17.97% mAP@0.5:0.95, outperforming the global-based DeCUR baseline by 1.55 and 0.86 percentage points respectively. These results confirm that preserving spatial structure while explicitly decoupling common and unique representations at the patch level leads to superior performance on downstream dense prediction tasks.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009.

[3] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Moco v2: Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[4] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2020.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

[6] Xiaoyang Huang, Hao Wang, Jie Zhang, Jing Zhang, and Dacheng Tao. On the comparison between multi-modal and single-modal contrastive learning. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[7] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baselines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1037–1045. IEEE, 2015.

[8] Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. Icafusion: Iterative cross-attention

guided feature fusion for multispectral object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[9] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[10] Yi Wang, Conrad M. Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decoupling common and unique representations for multimodal self-supervised learning (decur). In *Computer Vision – ECCV 2024, Lecture Notes in Computer Science, vol. 15087*. Springer, 2024.

[11] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[12] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 12310–12320, 2021.

[13] Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. Self-supervised multimodal learning: A survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.