Contra-Phy: Re-Introducing Boundary Awareness in Infrared Image Classification via Physics-Informed Contrastive Learning

Noyon Dey[0000−0001−5074−3281], Nasid Habib Barna[0009−0007−8953−1391], Jin Lu[0000−0003−1356−0202], and Suchendra M. Bhandarkar[0000−0003−2930−4190]

School of Computing, University of Georgia, Athens, GA 30602, USA
`{noyon.dey, nasidhabib.barna, jin.lu, suchi}@uga.edu`

Abstract. Infrared (IR) image classification is challenging due to limited visual detail, weak boundary cues, and foreground–background color consistency, which degrade discriminative information. These issues are particularly problematic in security and surveillance applications. Moreover, auxiliary information extracted from IR images is often unreliable, underscoring the need for more robust learning strategies that helps learn more usable object boundary cues and reliably improve classification performance under foreground–background consistency. To address this problem, we propose a physics-informed contrastive learning framework, ***Contra-Phy***. This framework incorporates physics-based cues (i.e., isotherms and temperature gradients) in a modified NT-Xent contrastive loss, to jointly capture semantic similarity and physical structure. These cues are applied only to positive pairs via ***positive masking***, preserving discriminative contrastive learning while promoting boundary-awareness. Extensive experiments on three datasets show consistent classification ***performance improvements***, and more ***controlled and spatially localized*** activations, indicating enhanced boundary information mining. Additionally, *Contra-Phy* contributes to improved robustness under stressed imaging test conditions. Beyond physics-based cues, we show that simple shape-based features such as Histogram of Oriented Gradients (HOG) can be seamlessly integrated into *Contra-Phy*. Unlike shape context (SC), this approach avoids explicit mask extraction (i.e., less computational overhead), thereby reducing dependency and reliability issues while improving performance through boundary mining.

Keywords: Infrared (IR) Image Classification · Foreground-background Consistency · Physics-informed · Contrastive Learning.

1 Introduction

Infrared (IR) image modality represents scenes through emitted thermal energy. This is in contrast to RGB (visible or color) image modality which reflects light for scene representations. Hence, IR image modality is particularly effective in low-light, nighttime, and adverse environmental conditions. For this reason, it is widely used in security and surveillance-related applications.

IR images are different than traditional RGB images as they do not have detailed visual information such as texture, edges, and color information. This significantly reduces visual details for any image classifiers, and in turn, makes it difficult for classification task.

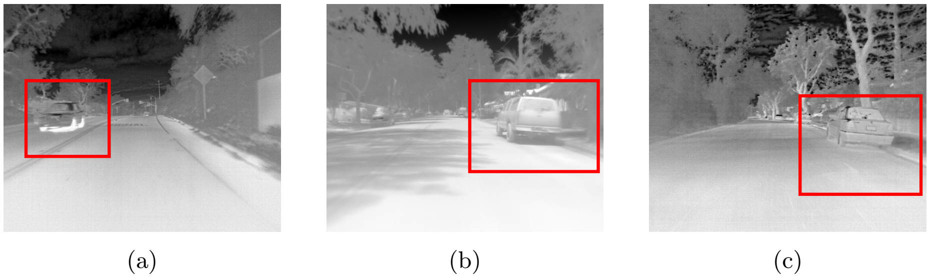


Fig. 1: Foreground–background color and appearance consistency in IR imagery (red-bordered regions), which weakens object boundary cues and often confuses the classifier, resulting in incorrect and unreliable classifications.

Apart from visual details, IR image also suffers other problems such as ***object boundary details, foreground-background consistency, auxiliary information visibility***, etc., which are crucial in classification task. Object boundaries help identify the shape cues for a specific class. Due to emissivity of certain parts, some of the important remaining parts may not be illuminated and hence, it looks like the background (non-illuminative parts) as illustrated in **Fig. 1**. This leads to a confusion, and potentially reduces classifier performances. Furthermore, it ***results in foreground-background color consistency*** problem where we do not have a clear distinction about the foreground (i.e., object of interest) and background (red bordered area in Fig. 1). Hence, the classifier gets confused and misclassify, which in turn, could play serious consequences in safety critical environments such as security and surveillance. To avoid such consequences, ***mining object boundaries, especially in foreground-background color consistent IR images is crucial***.

In some cases, a recent trend is to use auxiliary information such as shape context (SC)[2][1], and other attribute information[6], to guide during classification. However, extracting such auxiliary information is ***not always reliable in IR images*** (Fig. 5a). For example, to extract SC, we need to obtain a segmentation mask and then sample points from its contours. Hence, an accurate mask provides better shape information; however, reliable mask extraction is challenging in less visually rich modalities such as IR. This difficulty becomes more pronounced in noisy and degraded imaging conditions, such as those caused by adverse weather or environmental interference. On top of that, foreground-background color consistency also plays its part and makes it difficult to segment proper masks. For this reason, adding such auxiliary information in IR modality does not always guarantees improved performances. These problems require for a more robust approach that could reliably supplement such auxiliary information, and eliminate problems such as foreground-background color consistency yet reintroducing object boundary to improve classification performance.

To address this challenge, we propose ***Contra-Phy***, a physics-informed contrastive learning framework. This framework modifies the NT-Xent[4] contrastive loss by mining object boundary awareness through physics-based features. Specifically, we leverage IR-specific cues such as ***isotherms and temperature gradients*** within the contrastive objective. These cues represent subtle yet discriminative physical characteristics of IR imagery, and with such physics-based cues, the proposed framework helps reintroduce/mine usable object boundary signals that are often suppressed in IR images.

In addition, we investigate the integration of other informative cues, such as the Histogram of Oriented Gradients (HOG), within the ***Contra-Phy***. HOG provides explicit shape information that could be effectively leveraged for contrastive representation learning and performance improvement. Unlike more complex shape descriptors such as SC, HOG does not require explicit segmentation mask extraction. This property makes HOG-based features more reliable and avoids error propagation caused by inaccurate mask estimation. Furthermore, this demonstrates that the proposed *Contra-Phy* framework is flexible and can incorporate complementary cues beyond physics-based features. Consequently, integrating these cues improves overall classification performance. **In summary, our main contributions are:**

ⓐ We propose ***Contra-Phy***, a physics-informed contrastive framework that ***extract/re-introduce*** object boundaries in foreground-background color consistent IR images. Specifically, it integrates physics-based cues into the contrastive loss to enhance boundary awareness. To the best of our knowledge, this is the first work to incorporate such physics-based cues into a contrastive learning objective for IR image classification. ⓑ It mitigates foreground–background color consistency and improves performance across datasets under clean and stressed conditions, ***without*** stress-based augmentations. ⓒ It supports seamless integration of classical ***shape cues*** (e.g., HOG) ***without explicit mask extraction***, improving robustness and reliability. ⓓ We also identify consistent auxiliary phsyics-based cues that help improve performance.

2 Literature Review

Image classification in IR image modality is not widely explored whereas the RGB image modality has been greatly explored [7, 17, 23]. In addition, most of the foundation models in image classification are pretrained on RGB image-based large datasets which provides them thorough knowledge of RGB modality.

Lack of foundation models in IR: On the contrary, for the IR image modality, we do not have enough modality-specific resources such as foundation models. To the best of our knowledge, there lies only one IR image modality-specific foundation model InfMAE[15] that was pretrained on around 300k IR images.

This is again not directly comparable to the RGB pretrained models since they are trained on even bigger datasets (e.g., over a million images for ImageNet[5]).

Less Visual Details in IR: In addition, IR image modality does not provide detailed visual information (e.g., texture, edges, color, etc.) as in RGB modality. On top of that, we have other issues such as foreground-background consistency, and this greatly confuses the classification task. Classifier cannot identify accurate boundary information and hence confuses the overall classification task.

Contrastive learning approaches in IR: Contrastive learning has been a popular choice in classification and detection tasks, and IR modality has also been explored with this approach, especially in detection tasks[24]. Such is seen in[3] where a supervised contrastive learning was used to detect anomaly in IR images. Goyal et. al[10] used a self-supervised approach to detect and isolate hotspot from IR images. In another work, authors extend supervised learning in visible-infrared person identification[19], whereas in [26], authors used an unsupervised contrastive learning for the same task. In another work, authors used a supervised contrastive learning for hyperspectral image classification[12].

Physics-based Features: There are existing works on classification that address physics informed features such as temperature and emissivity[27]. The authors use attention-based mechanism and structured state-space model to jointly classify the IR-hyperspectral image. In another work[25], emissivity was used to detect objects. This state-space modeling is also used in conjunction with contrastive learning for IR image generation[14].

While contrastive learning has been explored in IR imagery, the use of physics-aware features within this framework remains largely unexplored (to the best of our knowledge). In this work, we leverage physics-informed contrastive learning to address foreground–background consistency and extract/re-introduce usable object boundaries for robust IR image classification.

3 Method

We propose ***Contra-Phy***, an enhanced physics-informed *NT-Xent*-based contrastive learning framework, that aims to reintroduce discriminative object boundary in IR images facing foreground-background color consistency problem (**Fig. 2**).

3.1 IR-specific Physics-based Features

Isotherm Histogram: Isotherm refers to a contour or region connecting points of equal temperature (or thermal intensity) in an IR image, characterizing constant-temperature structures in the scene [11],[8]. Given an input IR image tensor, we first convert it to a grayscale representation and linearly rescale the intensity values to the range $[0, 1]$. The grayscale image is then flattened into a one-dimensional vector, and its pixel intensities are quantized into K discrete bins, where $K = 16$ in our experiments. An isotherm histogram is constructed by accumulating the frequency of pixels falling into each intensity bin, followed by

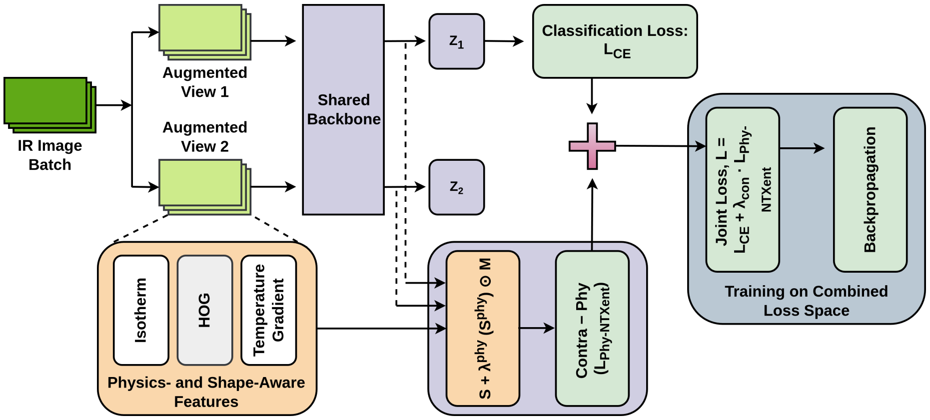


Fig. 2: Overview of the proposed ***Contra-Phy*** framework. An IR image batch is augmented into two views, encoded via a shared backbone (*ViT/InfMAE*), and trained with a joint objective combining cross-entropy (CE) and the proposed physics-informed contrastive loss ($L_{Phy-NTXent}$). Physics-based cues (isotherm histograms, temperature gradients) and shape-based cues (HOG) are integrated into the contrastive objective via ***positive masking***, selectively reinforcing similarity only for positive pairs. Positive masking was employed to mitigate confusions due to similar physical structures in negative samples.

ℓ_1 normalization. The resulting histogram provides a compact global representation of constant-temperature regions in the IR image, capturing the distribution of thermal intensity levels without requiring explicit object segmentation.

Temperature Gradient Magnitude: Temperature gradient denotes the spatial rate of change of thermal intensity in an IR image, highlighting local transitions and boundaries through variations in temperature across neighboring pixels [9], [11]. Given an input IR image tensor, we first convert it to a grayscale representation, and spatial temperature gradients are then computed using Scharr-like convolution kernels along the horizontal and vertical directions. The gradient magnitude at each spatial location is obtained as

$$\mathbf{M} = \sqrt{G_x^2 + G_y^2}, \tag{1}$$

where G_x and G_y denote the horizontal and vertical gradient responses, respectively. The resulting gradient magnitude map is min–max normalized to the range $[0, 1]$ to ensure scale consistency across samples. This representation captures local variations in thermal intensity and provides boundary-sensitive cues that are informative for IR image classification.

3.2 $\boldsymbol{Contra-Phy}(\mathcal{L}_{\textbf{Phy-NTXent}})$: Physics-Aware *NT-Xent* Loss.

Given a batch of paired augmented samples, we denote their learned embeddings as $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{B \times D}$, where B is the batch size. We first apply ℓ_2 normalization and construct the standard NT-Xent similarity matrix using cosine similarity scaled by a temperature parameter τ (learnable):

$$\mathbf{S} = \mathbf{S}_{ij} = \frac{\mathbf{z}_i^\top \mathbf{z}_j}{\tau}. \tag{2}$$

To incorporate physics-informed cues, we additionally extract physics-based cues from each sample. Let $\mathbf{g}$ be normalized physics feature. Then, we compute pairwise cosine similarity matrices:

$$\mathbf{S}^{\text{phy}} = \mathbf{g}\mathbf{g}^\top \tag{3}$$

We introduce a physics-aware similarity bonus that is applied *only* to positive pairs through **positive masking (M)**:

$$\mathbf{sim^{phy}} = \mathbf{S} + \lambda^{\text{phy}} \left(\mathbf{S}^{\text{phy}}\right) \odot \mathbf{M}, \tag{4}$$

where λ^{phy} controls the strength of the physics regularization (1 in our case), $\mathbf{M}$ is a binary mask indicating positive pairs, and $\odot$ denotes element-wise multiplication.

Finally, the physics-informed ***NT-Xent*** loss is computed using a softmax cross-entropy objective:

$$Contra - Phy(\mathcal{L}_{\text{Phy-NTXent}}) = -\log \frac{\exp\left(\text{sim}^{\text{phy}}(\mathbf{z}_i, \mathbf{z}_j)/\tau\right)}{\sum_{k=1}^{2B} 1_{[k \neq i]} \exp\left(\text{sim}^{\text{phy}}(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)} \tag{5}$$

This formulation encourages embedding similarity to align not only in representation space but also with underlying thermal structure and physical cues.

Positive masking in *Contra-Phy*: To avoid distorting the global contrastive structure, the physics-informed similarity is applied exclusively to positive pairs. Specifically, we construct a binary positive-pair mask $\mathbf{M} \in \{0,1\}^{2B \times 2B}$, where $\mathbf{M}_{ij} = 1$ if and only if samples i and j form a designated positive pair (i.e., i and j correspond to the two augmented views of the same instance, where one view is in the first batch and the other in the second batch), and $\mathbf{M}_{ij} = 0$ otherwise. For each sample, there exists exactly one positive pair.

The physics-based similarity terms are explicitly prevented from influencing negative pairs, ensuring that inter-instance separation is governed solely by the learned representation space. This design avoids boundary collapse and suppresses misleading thermal similarities arising from foreground–background intensity consistency. As a result, the selective masking strategy preserves the discriminative behavior of NT-Xent while encouraging physics-consistent alignment between positive pairs.

3.3 Overall Joint Objective

The training objective consists of two components: a supervised CE classification loss ($\mathcal{L}_{\text{CE}}$) and ***Contra-Phy*** loss: a physics-informed contrastive loss ($\mathcal{L}_{\text{Phy-NTXent}}$). The final training loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{con}} \cdot \mathcal{L}_{\text{Phy-NTXent}}, \tag{6}$$

where λ_{con} controls the contribution of the contrastive loss and is gradually increased during the initial training phase until it reaches a fixed value of one, where it remains for the rest of training.

4 Datasets and Baselines

4.1 Datasets

In this work, we used three IR image datasets: (a) **FLIR+M3FD** ($\approx 10k$ samples with 50:50 train-test split), (b) **LTDv2** ($\approx 16k$ samples with a 90:10 train-test split), and (c) **Stanford cars** (IR version: $\approx 16.2k$ samples with a 50:50 train-test split). Owing to the limited availability of labeled IR image classification datasets, we curated and manually annotated subsets from existing benchmark IR datasets to support supervised classification.

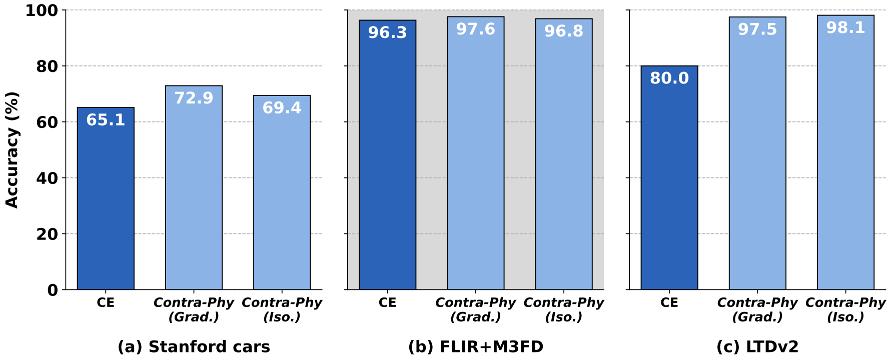
For the FLIR+M3FD dataset, we used two benchmark IR dataset named the FLIR[22] and M3FD[16], and assembled $\approx 10k$ images for two classes. For the LTDv2[18] dataset, we assembled $\approx 16k$ images (for two classes) from 1 million frames ranging throughout the year. For the Stanford Cars dataset [13] (196 classes), we construct a synthetic IR dataset by translating RGB images into the IR domain using a fine-tuned AttentionGAN [21]. The AttentionGAN model was fine-tuned on the paired FLIR dataset and subsequently applied to the RGB images from Stanford Cars to generate their corresponding IR representations.

4.2 Baselines

(a) **CE only:** We also compare the *Contra-Phy* with CE only classification loss.
(b) **Clean-Contrastive learning:** We employ the ***NT-Xent*** loss with ***positive masking*** and ***no physics-informed*** features.
(c) **No-masking:** In this scenario, we use the same physics-based features with contrastive learning, however, with no positive masking.
(d) **Feature fusion:** We adopt a feature-level fusion strategy where visual representations from the backbone are combined with physics-based cues. Specifically, the *ViT* extracts a global image embedding from the [CLS] token, while the physics-based cues are flattened and projected into a compact embedding via a lightweight MLP. These two representations are concatenated to form a fused feature vector, which is then used for both classification.
(e) **Input fusion:** We extend a pre-trained *ViT* to accept 4-channel input by adapting the patch embedding layer, where the additional channel encodes physics-based cues. The resulting [CLS] representation is used for both classification.

5 Result and Discussion

5.1 Experimental Results



(a) Stanford cars (b) FLIR+M3FD (c) LTDv2

Fig. 3: ***Contra-Phy*** improves classification performance across datasets. Experimental results demonstrate consistent performance gains across datasets when ***ViT*** is trained along with ***Contra-Phy*** loss. Here, CE denotes cross-entropy loss, while Grad.≈Temparature Gradient, and Iso.≈Isotherm.

We conduct experiments on three IR datasets, as described in Section 4.1, using two physics-based features: isotherms and temperature gradients. The corresponding results are presented in **Fig. 3**. As shown, integrating physics-informed features within a contrastive learning framework consistently improves overall classification performance across all datasets than the CE only model. These results demonstrate the effectiveness and robustness of incorporating physics-aware cues for enhancing IR image classification performance.

Why positive masking? Fig. 4 validates the effectiveness of the proposed ***positive masking*** strategy in ***Contra-Phy***. It consistently improves classification performance over the baselines while promoting more controlled boundary awareness (**Fig. 6**), motivating its adoption in the final framework.

5.2 Limitations of Auxiliary Data Fusion in IR

The IR modality presents challenges for reliable auxiliary cue extraction due to limited visual detail and issues such as foreground–background consistency, which worsen under visual degradations. In such conditions, auxiliary information obtained ***via detectors or preprocessing is often unreliable***, and may degrade rather than improve classification performance (**Fig. 5a**).

Fig. 5a shows results on a subset of the vehicle dataset where semantic attributes and SC are incorporated via **late fusion** at the ***ViT*** classification layer (without contrastive loss). Contrary to expectations, combining these auxiliary

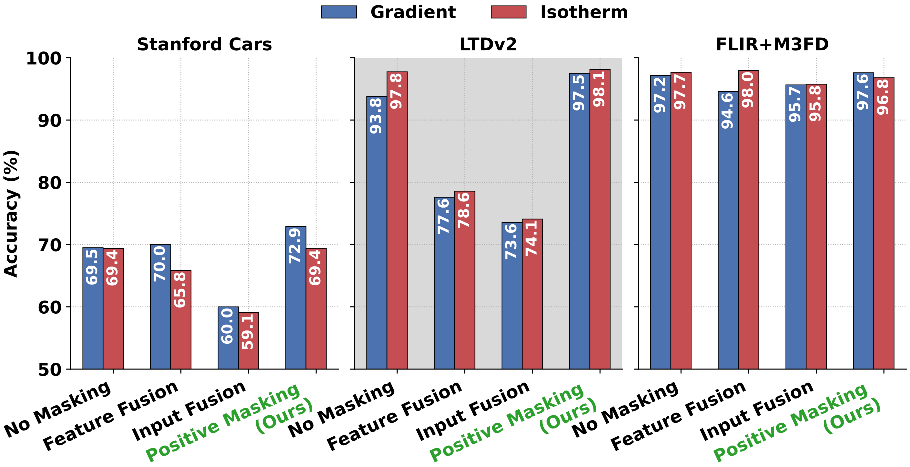


Fig. 4: ***Contra-Phy*** (i.e., Positive Masking (Ours)) improves classification performance across datasets compared to other baselines. Experimental results with the ***ViT*** model on (a) Stanford Cars, (b) FLIR+M3FD, and (c) LTDv2 demonstrate consistent performance gains with ***Contra-Phy***. It further achieves more controlled and localized activation than the baselines, which we show in Fig. 6.

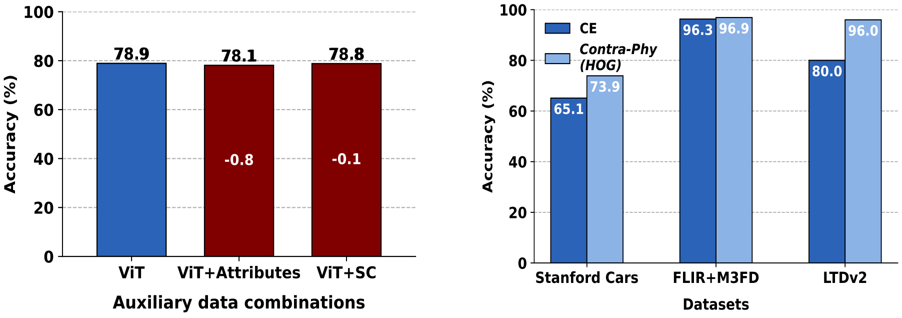
cues leads to a ***slight performance degradation***, indicating that they are not reliably informative and may ***introduce noise rather than complementary signal*** in the IR modality. This observation highlights the need for a more principled integration strategy. In light of this, the proposed physics-aware contrastive learning framework enforces physically meaningful consistency during representation learning, enabling more effective utilization of auxiliary cues and resulting in consistent performance improvements (**Fig. 3**).

5.3 Incorporating Shape Cues via *Contra-Phy*

Furthermore, SC computation depends on accurate object masks and contour sampling, making it unreliable in IR imagery due to low contrast, limited texture, and foreground–background consistency. Consequently, imperfect masks often produce noisy SC features that can degrade classification performance (Fig. 5a).

To mitigate this, we adopt a more robust shape descriptor within a contrastive framework. Specifically, we incorporate Histogram of Oriented Gradients (HOG) features into the ***Contra-Phy*** loss (Section 3.3), ***avoiding explicit mask dependency*** while preserving meaningful structural cues for IR images.

We present this experiment's result in Fig. 5b. We can see that classic and simple HOG-based features with contrastive learning indeed provided similar overall classification performance improvements. Consequently, this demonstrates that even simple and classical feature descriptors, integrated within the



(a) (b)

Fig. 5: (a) Unreliable explicit auxiliary cue extraction (e.g., attributes and SC) in IR and incorporating them does not consistently improve classification performance. (b) In contrast, integrating HOG features within the proposed ***Contra-Phy*** framework yields substantial and consistent performance improvements.

Contra-Phy contrastive learning framework, can effectively enhance representation learning and lead to improved classification performance.

5.4 Ablation Studies on Shape and Physics Cues

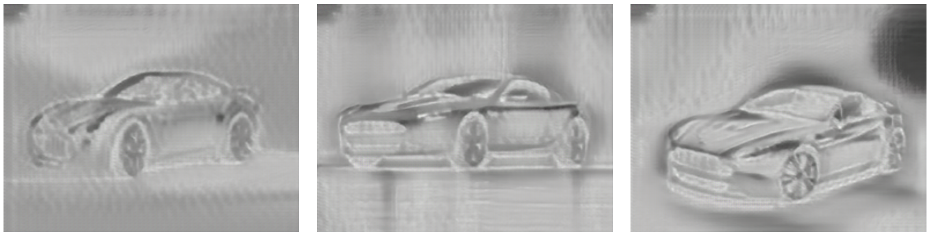
We further conduct experiments combining physics-based and shape-based cues to examine their joint impact on overall classification performance. Specifically, we conduct experiments using ***both IR- and RGB-pretrained*** foundation models. We investigate ViT[7] and ***InfMAE***[15], which are fine-tuned and trained on our IR datasets.

The ablation results, summarized in **Table 1**, demonstrate that integrating these complementary cues within the proposed ***Contra-Phy*** contrastive learning framework generally improves performance across different physics–shape combinations. Specifically, ***CE+Grad. (i.e., CE loss with Contra-Phy with temperature gradient cue) combination remains consistent*** across all the datasets. On the other hand, other cues and their combinations also improves performances, however, they are not consistent across all the datasets.

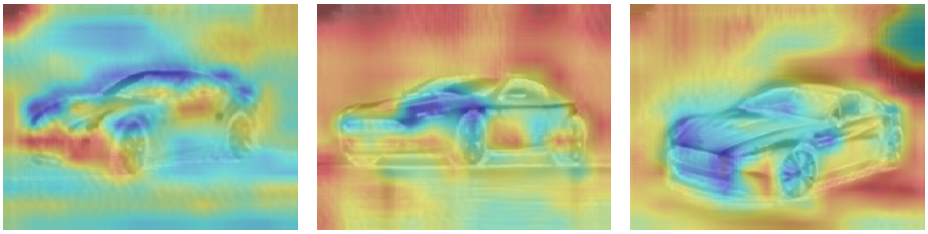
Most configurations outperform their respective baselines, with only a few exceptions (three on FLIR+M3FD and one on LTDv2 with InfMAE), likely due to modality-specific noise, limited discriminative structure, or mild overregularization. Overall, these results show that integrating physics-informed and shape-based cues within the contrastive ***Contra-Phy*** framework consistently improves classification performance.

5.5 *Contra-Phy* Improves Boundary Awareness

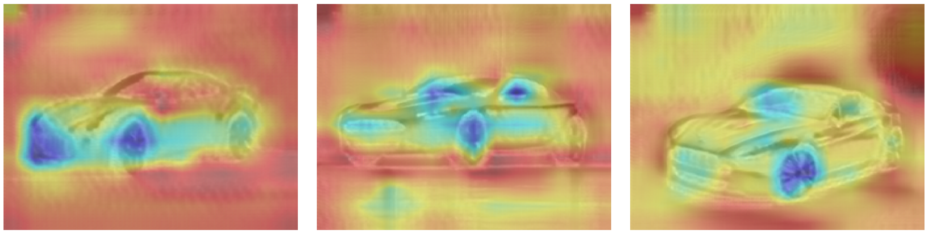
Incorporating physics-informed cues into the ***Contra-Phy*** leads to substantially improved activation maps (i.e., better object boundary delineation)(**Fig. 6**).



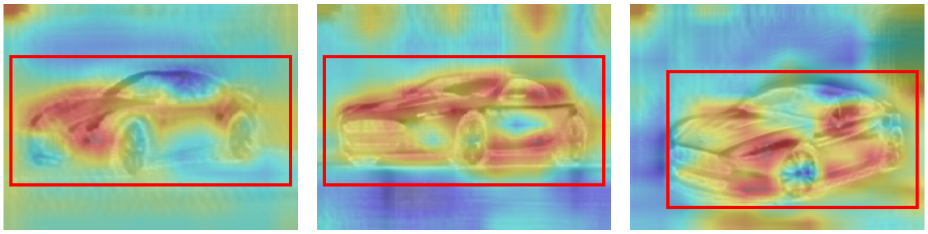
Row 1: Testing IR image samples.



Row 2: CE-only training results in diffuse and poorly localized activation maps.



Row 3: CE + Clean Contrastive learning, without physics-informed cues, yields less controlled and weakly localized activation maps.



Row 4: ***Contra-Phy*** (with Grad.) physics-informed contrastive learning produces more accurate and spatially focused activation maps aligned with the object of interest (red-bordered).

Fig. 6: Grad-CAM[20] activation map comparison across training objectives. Physics-informed (i.e., temperature gradient) contrastive learning (Row 4) produces more ***spatially focused and controlled activations***, aligned with the object of interest, whereas CE-only training (Row 2) and CE + Clean Contrastive learning (Row 3, ***without physics-informed*** cues) ***exhibit diffuse and background-dominated*** activations. The improved localization in Row 4 indicates that ***Contra-Phy*** effectively ***reintroduces*** usable object boundary, leading to discriminative representation learning and improved classification.

Table 1: Ablation studies on physics-aware and shape-based cues. Results show consistent improvements over the baseline across all datasets using ***ViT*** and ***InfMAE***. CE denotes cross-entropy loss, Grad. denotes temperature gradient features, and Iso. denotes isotherm features. **x** denotes the best improved result, and **x** denotes improved results than (CE and CE+Clean Contrastive).

Experiments	**FLIR+M3FD**	**LTDv2**	**Stanford Cars**
	Accuracy (%)	**Accuracy (%)**	**Accuracy (%)**
***ViT* Experiments**			
CE only	96.3	80.0	65.1
CE + Clean Contrastive	96.8	97.5	72.16
CE + Grad.	**97.6**	**97.5**	**72.9**
CE + Iso.	**96.8**	**98.1**	69.4
CE + HOG	**96.9**	96	**73.9**
CE with HOG+Iso.	**97.7**	**98.4**	**73.6**
CE with HOG+Grad.	**96.8**	**98.0**	**73.15**
CE with Iso.+Grad.	**97.3**	**98.4**	**73.3**
CE with Iso.+Grad.+HOG	**96.9**	**98.15**	**73.64**
***InfMAE* Experiments**			
CE only	96.7	81.2	50.3
CE + Clean Contrastive	92.3	95.8	54.9
CE + Grad.	**98.1**	**98**	**55.3**
CE + Iso.	96.1	**96.0**	**55.5**
CE + HOG	95.6	**96.1**	**56.0**
CE with HOG+Iso.	**97.5**	93.56	52.95
CE with HOG+Grad.	**97.85**	82.2	**55.2**
CE with Iso.+Grad.	**97.44**	72.0	54.52
CE with Iso.+Grad.+HOG	96.11	95.0	50.5

Consequently, the learned representations exhibit improved spatial coherence, leading to activation maps that are sharper, more localized, and less influenced by background clutter than the CE-only and CE + Clean Contrastive baselines.

5.6 Discussion and Implication

Consistent Gains Across Models: We evaluate ***Contra-Phy*** on *ViT* (RGB) and *InfMAE* (IR), where it consistently outperforms the baselines. Notably, incorporating ***temperature gradient*** cues with CE ***yields stable gains*** **(Table 1)** across both architectures, while combining multiple physics- and shape-based cues further improves robustness across modalities.

Improved Robustness to False Positives: Beyond improving classification accuracy, *Contra-Phy* produces more controlled and spatially focused activations during inference (Fig. 6). These activation patterns indicate that models trained with *Contra-Phy* rely more consistently on the true object of interest rather

Table 2: Ablation studies using ***ViT*** model on physics-aware and shape-based cues ***under stressed testing conditions***. Stresses are applied ***only at test time***, with ***no corresponding perturbations introduced during training***. Results show consistent performance improvements over baseline methods across all datasets. Performance gains over **both baselines** (i.e., CE only, and CE + Clean Contrastive) are highlighted in underlined green (e.g., **x**), improvements over only one baseline are shown in bold (e.g., **x**).

Experiments	Clean	Noise	Blur	Illumination	Occlusion	Atmospheric	Adversarial
FLIR+M3FD							
CE only	96.3	84.94	97.36	96.6	94.31	97.46	94.08
CE + Clean Contrastive	96.83	90.44	96.65	96.85	93.44	96.5	96.11
CE + Grad.	**97.58**	**94.25**	**97.48**	**97.71**	**96.67**	**97.77**	**97.06**
CE + Iso.	**96.85**	**94.15**	**97.56**	**97.06**	**95.02**	**97.11**	**96.42**
CE + HOG	**96.9**	**92.79**	**96.81**	**97.06**	**94.33**	**96.77**	**95.96**
LTDv2							
CE only	80	78.39	57.44	78.45	75.08	69.13	79.91
CE + Clean Contrastive	97.95	97.62	98.15	96.36	93.26	94.58	98.08
CE + Grad.	**97.49**	**91.34**	**90.02**	**93.39**	**91.41**	**91.34**	**97.69**
CE + Iso.	**98.08**	**96.23**	**98.22**	**97.22**	**93.92**	**98.28**	**98.08**
CE + HOG	**96.03**	**92.47**	**95.97**	**92.73**	**90.81**	**92.66**	**96.1**
Stanford Cars							
CE only	65	28.88	47.41	63.29	45.18	55.13	54.81
CE + Clean Contrastive	72.16	39.3	45.68	69.48	54.91	56.61	67.29
CE + Grad.	**72.91**	**39.21**	**47.54**	**70.77**	**55.07**	**58.29**	**67.4**
CE + Iso.	**69.42**	**32.32**	44.85	**67.6**	**51.26**	**55.44**	**63.52**
CE + HOG	**74**	**42.89**	**48.82**	**71**	**56.4**	**59.02**	**69.58**

than on spurious background regions. As a result, the proposed approach is less prone to false positive predictions compared to models trained without physics-informed guidance.

Expected Benefits in Challenging Imaging Scenarios: By promoting boundary-aware and spatially focused representations (**Fig. 6**), ***Contra-Phy*** is expected to improve robustness under stressed conditions (e.g., camouflage, distortion). We further validate this using stressed test samples generated via augmentations (***without applying them during training***), including noise, blur, illumination, occlusion, atmospheric, and adversarial stress. As shown in **Table 2**, ***Contra-Phy*** improves performance in the majority of cases despite these perturbations.

Limitations and future work: Large-scale, publicly available IR image classification datasets remain limited compared to their RGB counterparts. As a result, our experiments are conducted on three public IR datasets that were curated for classification through manual annotation. While this enables controlled evaluation, the curated and translated datasets represent a necessary and early step toward establishing standardized benchmarks for IR image classification, while also providing a practical foundation for systematic study in the absence of large-scale annotated IR datasets. Future work includes the development of larger, diverse IR image classification datasets and more extensive evaluations to further explore and validate informative cues in IR imagery. Furthermore, although ***Contra-Phy*** effectively integrates HOG-based cues, exploring additional lightweight, mask-free features without explicit preprocessing is a promising direction for improving flexibility and performance.

6 Conclusion

We propose ***Contra-Phy***, a framework that integrates physics-informed cues (isotherm and temperature gradients) to address foreground–background consistency in IR imagery and improve classification. The method enhances performance while yielding more controlled (i.e., boundary aware) and interpretable activations, with consistent gains across three datasets. In particular, temperature gradient cues provide stable improvements, and the framework remains robust under stressed conditions. Additionally, *Contra-Phy* supports seamless integration of shape-based HOG features, avoiding explicit preprocessing and potential extraction errors.

References

1. Basu, S., Dey, N., Bhandarkar, S.M., Wolbach, S.: Integrating shape-and cnn-based features for zero-shot and low-shot learning (2024)
2. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE transactions on pattern analysis and machine intelligence **24**(4), 509–522 (2002)
3. Bommes, L., Hoffmann, M., Buerhop-Lutz, C., Pickel, T., Hauch, J., Brabec, C., Maier, A., Marius Peters, I.: Anomaly detection in ir images of pv modules using supervised contrastive learning. Progress in Photovoltaics: Research and Applications **30**(6), 597–614 (2022)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Dey, N., Bhandarkar, S.M.: Exploratory insights into late-fused attribute cues for rgb-based military vehicle recognition. In: Advances in Visual Computing. pp. 454–467. Springer Nature Switzerland, Cham (2026)
7. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Gade, R., Moeslund, T.B.: Thermal cameras and applications: a survey. Machine vision and applications **25**(1), 245–262 (2014)
9. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Prentice Hall (2009)

10. Goyal, S., Rajapakse, J.C.: Self-supervised learning for hotspot detection and isolation from thermal images. Expert Systems with Applications **237**, 121566 (2024)
11. Holst, G.C.: Common Sense Approach to Thermal Imaging. SPIE Press (2000)
12. Huang, L., Chen, Y., He, X., Ghamisi, P.: Supervised contrastive learning-based classification for hyperspectral image. Remote Sensing **14**(21), 5530 (2022)
13. Krause, J., Deng, J., Stark, M., Fei-Fei, L.: Collecting a large-scale dataset of fine-grained cars (2013)
14. Li, B., Ma, D., He, F., Zhang, Z., Zhang, D., Li, S.: Infrared image generation based on visual state space and contrastive learning. Remote Sensing **16**(20), 3817 (2024)
15. Liu, F., Gao, C., Zhang, Y., Guo, J., Wang, J., Meng, D.: Infmae: A foundation model in the infrared modality. In: European Conference on Computer Vision. pp. 420–437. Springer (2024)
16. 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)
17. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
18. Parola, M., Aakerberg, A., Johansen, A.S., Nikolov, I.A., Cimino, M.G., Nasrollahi, K., Moeslund, T.B.: Ltdv2: A large-scale long-term thermal drift dataset for robust multi-object detection in surveillance. Authorea Preprints (2025)
19. Qian, Y., Tang, S.K.: Multi-scale contrastive learning with hierarchical knowledge synergy for visible-infrared person re-identification. Sensors (Basel, Switzerland) **25**(1), 192 (2025)
20. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
21. Tang, H., Liu, H., Xu, D., Torr, P.H., Sebe, N.: Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. IEEE transactions on neural networks and learning systems **34**(4), 1972–1987 (2021)
22. TELEDYNE: Teledyne flir dataset. https://oem.flir.com/solutions/automotive/adas-dataset-form/ (2025), accessed: 2025-12-18
23. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
24. 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. Frontiers in Physics **11**, 1193245 (2023)
25. Yang, F., Qiu, H., Cheng, I.: Emissivity-guided thermal image encoding: A robust multitask learning approach for object detection. In: 2025 IEEE International Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 000–000. IEEE (2025). https://doi.org/10.1109/MIPR67560.2025.00089
26. Zhang, Y., Zhang, C., Ma, H., Li, Z., Wang, Z., Wei, C.: Unsupervised infrared–visible person re-identification by multi-level dual-stream contrastive learning. Neurocomputing **634**, 129895 (2025)
27. Zhao, E., Qu, N., Wang, Y., Gao, C., Zeng, J.: Tebs: Temperature–emissivity–driven band selection for thermal infrared hyperspectral image classification with structured state-space model and gated attention. International Journal of Applied Earth Observation and Geoinformation **142**, 104710 (2025)