DD-LIVM: Pioneering Cross-Domain Photovoltaic Defect Detection Using Large Infrared-Visible Model

Yinan Zhu[1], Meng Xue[1], Haiyan Hu[1], Cong Zhang[2], Xiaoyi Fan[2], Qian Zhang[1]
[1]The Hong Kong University of Science and Technology [2]Jiangxing Intelligence Inc.

Abstract

Photovoltaic (PV) defect detection is crucial for preventing power efficiency loss and fire hazards. The industry primarily relies on the fusion of infrared and visible images for defect localization and diagnosis. However, current detection methods exhibit poor generalizability in new site environments or with altered imaging setups. While recent infrared and vision foundation models (FM) facilitate domain-invariant feature maps extraction, directly concatenating them and fine-tuning achieves limited generalizability gain to PV defect detection, due to the asymmetric dual-modal semantics of defects. In this paper, we present the first large infrared-visible model *DD-LIVM* to enable cross-domain defect detection. The key innovation of DD-LIVM lies in its defect-specific three-step fine-tuning strategy, which utilizes alternating modality masking. Prior to feature fusion and joint fine-tuning, the infrared and visible FM encoders are alternately masked and optimized to enhance their individual semantic utility for defect localization visibility and classification granularity, with feature distances among different defect types regulated through contrastive learning. This approach allows for the extraction of generalizable and defect-specific feature maps. Moreover, for practical employment of DD-LIVM, we propose a domain-agnostic spatial alignment algorithm for infrared-visible images before dual-modal fusion, and develop source data augmentation and adaptive detection head selection schemes based on defects' infrared characteristics to further enhance the generalizability. Extensive experiments on 7,078 dual-modal images from 9 real-world scenarios across 4 cities' PV stations demonstrate that DD-LIVM achieves an accuracy of 87.7% for cross-domain defect detection, surpassing state-of-the-art methods by 17.3%.

ACM MOBICOM '25, Hong Kong, China

ACM ISBN 979-8-4007-1129-9/25/11
https://doi.org/10.1145/3680207.3765266

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Computer systems organization** → *Embedded systems.*

Keywords

Photovoltaic Panel Defect Detection, Large Infrared-Visible Model, Alternating Modality Mask

ACM Reference Format:
Yinan Zhu[1], Meng Xue[1], Haiyan Hu[1], Cong Zhang[2], Xiaoyi Fan[2], Qian Zhang[1]. 2025. DD-LIVM: Pioneering Cross-Domain Photovoltaic Defect Detection Using Large Infrared-Visible Model. In *The 31st Annual International Conference on Mobile Computing and Networking (ACM MOBICOM '25), November 4–8, 2025, Hong Kong, China.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3680207.3765266

1 Introduction

As critical infrastructure for renewable energy, the global deployment of photovoltaic (PV) systems has significantly reduced fossil fuel dependency and accelerated the green energy transition [31, 33, 37]. Regular defect detection of PV panels is essential, as undiagnosed issues-such as grass occlusion or circuit failures-can severely degrade power generation efficiency, compromise grid stability, and even pose fire hazards [1, 18, 41]. Currently, the mainstream industrial solutions employ autonomous drone inspection technology, which fuses the drone-captured **thermal infrared images and visible light images** to automatically locate the PV defects and diagnose the defect types [9, 19], as shown in Figure 1.

However, existing PV defect detection methods [5, 9, 19, 35, 48] often produce models whose extracted defect features are entangled with domain-dependent contexts. This reliance on specific data distributions results in **low generalizability** when faced with data heterogeneity caused by new environmental conditions (*e.g.*, terrain backgrounds, solar irradiance levels) or altered imaging setups (*e.g.*, camera intrinsics and extrinsics, drone flight altitudes). This limitation forces repetitive labor-intensive data re-collections and manual defect re-annotations for constant model fine-tuning, failing to enable cross-domain defect detection. Current domain generalization methods [4, 15, 23, 45] also have poor performance under the above complex and variable domain shift factors, such as frequent changes in solar irradiance

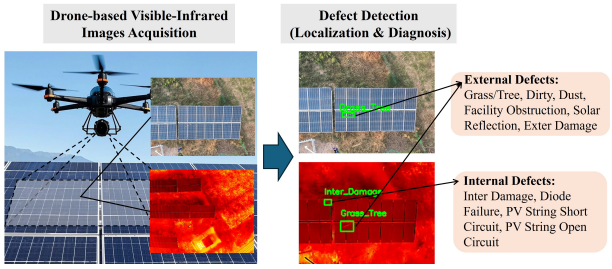

Figure 1: Illustration of PV Defect Detection Scenario

or ambient temperatures. In recent years, the emergence of infrared foundation models (FMs) has facilitated the extraction of domain-invariant infrared feature maps [26]. When combined with vision FMs, this presents an opportunity to enhance the generalizability of PV defect detection. Despite this, directly concatenating them and jointly fine-tuning cannot achieve satisfied generalization performance. This is because the infrared and visible semantics of PV defects have independent physical mechanisms (*i.e.*, electrothermal characteristics and appearance morphology, respectively) and are not a simply complementary relationship [18], which can lead to feature misalignment and model overfitting to source training data. For example, diode malfunctions incur hotspots in infrared images without any cues in visible images, while the visible encoder may excessively learn the location contextual information of diode malfunctions in the training images during intuitive fine-tuning, yielding source-domain overfitting.

To fill this gap, we dive into the dual-modal semantics difference of internal and external defects, and observe that 1) infrared images can capture defect-induced hotspots with all defects' location information, but only four low-resolution shapes, and thus fail to distinguish the specific forms of internal and external defects; 2) visible images contain only external defects' location information and appearance features such as grass obstruction, while internal defects are totally invisible. Consequently, these two modalities serve distinct semantic purposes regarding localization visibility and classification granularity for different types of defects. Thus, *the primary challenge during fine-tuning is optimizing the dual-modality encoders to leverage their separate semantic utilities*, thereby mitigating the risks of source-domain overfitting.

In this paper, we present *DD-LIVM*, the first large infrared-visible model to enable precise cross-domain detection for ten common types of PV defects without any prior information of new domains. Our detection framework merges the latest FM encoders of infrared and visible modalities to extract feature maps, and incorporates a novel contrastive learning head for regulating feature gaps of different defects besides standard detection heads. To address the above challenge associated with defect semantics, we innovatively develop ***a defect-specific three-step fine-tuning strategy (DTFT) for the dual-modal pretrained FM encoders using alternating modality masks***. The process is as follows: First, we mask the visible module and fine-tune the infrared encoder to focus on hotspot localization and categorize the ten types of defects into four classes based on hotspot shapes. The feature gaps of distinct-shape defects are enlarged while gaps of same-shape defects are reduced, through contrastive learning. Second, with the infrared module masked, we fine-tune the visible encoder to concentrate solely on detecting visible external defects. The internal defect features are grouped together with increased distance from external defect features. Finally, we jointly fine-tune both encoders without any modality masks on the all-type defect detection task, aligning the location information of external defects and fusing dual-modal incomplete defect classification information for finer granularity. Through this iterative three-step fine-tuning, we achieve semantic-level dual-modal fusion, resulting in generalizable and defect-specific infrared-visible feature maps for cross-domain defect detection.

Moreover, we propose two further designs for the practical employment of DD-LIVM. Firstly, before dual-modal fusion, the collected infrared and visible images should be spatially aligned. Since the view and position offsets of new domains may totally differ from the source domain due to varying camera parameters or drone altitudes and have no prior information, spatial alignment is hard to achieve. To this end, we propose a ***universal spatial alignment algorithm for dual-modal images*** based on the consistency of two images' PV panel widths and relative positions, by extracting the contours of PV panels after background removal. Secondly, due to the tiny size and irregular shapes, infrared hotspot detection may still be affected by variant domain shifts. To overcome this, we leverage the hotspot shape and temperature distribution ***(HSTD) information for source data augmentation and adaptive detection head selection***. The former simulates image transformation according to HSTD changes with environments, thereby increasing the source-domain diversity of hotspot forms. The latter adds HSTD similarity-based defect sample perturbation to target images to select the detection head with high robustness.

We implemented DD-LIVM with over 400 million parameters and collected a dataset of 7,078 infrared and visible images featuring 7,297 defects across 10 defect types from real-world PV power stations in four cities, encompassing a total of nine different settings for evaluating generalizability. This dataset covers a broad spectrum of domain shift factors - site terrain backgrounds, camera devices and positions, drone flight heights and views, ambient temperature and solar irradiance. Experimental results demonstrate that the average accuracy of DD-LIVM can reach 87.7% in cross-domain defect detection, outperforming the SOTA schemes

by 17.3%. Moreover, DD-LIVM can achieve above 80% detection accuracy under all nine settings.

In summary, our main contributions are as follows.

- To the best of our knowledge, we introduce the first large infrared-visible model DD-LIVM that enables cross-domain PV defect detection. DD-LIVM can generalize to varying site environments and imaging setups without requiring any prior knowledge.
- We present a defect-specific three-step fine-tuning strategy to optimize the infrared and vision FM encoders for semantic-level fusion regarding defect localization visibility and classification granularity, based on alternating modality masks. Consequently, generalizable and defect-specific feature maps are obtained in DD-LIVM.
- For practical employment of DD-LIVM, we propose a domain-agnostic spatial alignment algorithm for dual-modal images based on the consistency of PV panel widths and relative positions. Besides, we develop HSTD-based source data augmentation and adaptive head selection schemes to enhance tiny hotspot detection across domains.
- We collect a dataset of 7,078 dual-modal images with 7,297 defects in real-world PV stations under nine scenarios with various environmental and imaging conditions. Our experiments over it verify DD-LIVM's superior generalizability in defect detection. The dataset samples of dual-modal images are publicly released in [8].

2 Background

Before elaborating on DD-LIVM's technical details, we introduce some backgrounds of this work.

PV Panel Defects. PV panel defects with 10 common types in Figure 1 can arise from diverse physical factors [41]. For example, vegetation and facility obstruction caused by nearby trees or buildings create local current mismatches, which are visible on the PV panel and also trigger thermal stress and hotspots. Electrical failures like string short circuits from insulation breakdown or open circuits due to connector corrosion or solder fatigue, are completely invisible and induce hotspots as well. Hence, both infrared and visible images are required to detect these defects.

Defect Detection. As a similar task to object detection, defect detection follows the backbone-neck-head framework [24, 36] as well, where the backbone extracts the infrared-visible feature maps, the neck conducts feature aggregation or fusion, and the head utilizes the fused features to predict the defect locations and types. So, the backbone is critical for obtaining generic defect features.

Current backbones perform poor generalization capability [19, 20, 38, 39] due to the entanglement of defect features with domain environmental contexts when simply using YOLO [20, 50], CNNs [28, 39] or shallow vision transformers [3, 38] for feature extraction. Accordingly, current detection methods cannot apply to new domains with complex shifts (see Figure 2). Additionally, for current dual-modal object detection, even the SOTA solutions GM-DETR [47] and DPDETR [11], still focus on dual-modal better fusion on a single dataset and their backbones fail to extract generic features of defects as well.

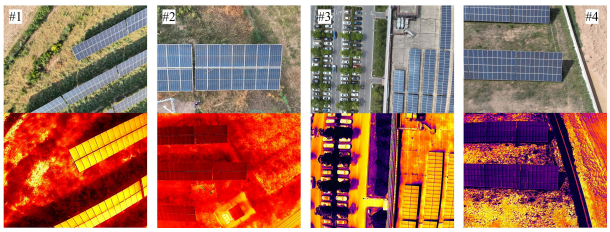

Figure 2: Collected visible and infrared images in *different domains* (#1~#4). Compared with #1, #2 differs in camera position and flight altitude, #3 evidently differs in terrain backgrounds (plain to rooftop), and #4 additionally differs in solar irradiance.

Infrared and Vision FMs. Recent years have emerged the first infrared FM, InfMAE [26], which is trained with large-scale 305K infrared images in different environments and demonstrates powerful learning capabilities of general representations with designed three multi-scale encoder layers. Besides, among all popular vision FMs [2, 30, 46], the SOTA one for object detection generalization, *i.e.*, FM-FSOD [14] based on DINOv2 structures [34] is also proved to surpass detection-specific methods by integrating with existing detection heads. Thus, we can leverage the two FM encoders as backbones to extract infrared and visible features, respectively. *Note that, currently there is no existing infrared-visible multimodal FM.*

Challenges in Applying FMs. We can concatenate the feature maps from the two modalities' FM encoders as the backbone output for defect detection. Since defect detection has unique physical representations that totally differ from object detection in autonomous driving or remote sensing fields, it is necessary to fine-tune the FM encoders together with the detection heads.

However, direct concatenating and fine-tuning them will lead to feature misalignment and model overfitting to source-domain training images, because infrared-visible semantics are not simply complementary and one modal's encoder may excessively learn semantically irrelevant information from the other modal. Therefore, the dual-modal FM encoders cannot be intuitively fine-tuned to the defect detection field for generalizable feature extraction.

3 System Design

In this section, we will introduce the detailed design of DD-LIVM. Figure 3 shows the working pipeline and framework

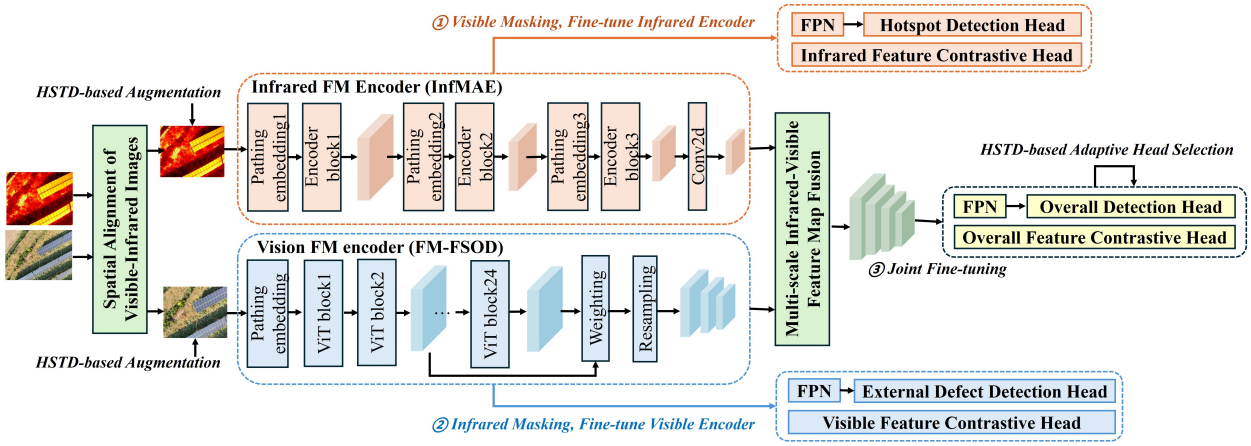

Figure 3: The working pipeline and framework of DD-LIVM, consists of three parts: 1) spatial alignment and preprocessing, 2) defect-specific three-step fine-tuning on backbones and multi-scale feature fusion, 3) HSTD-based data augmentation and head selection.

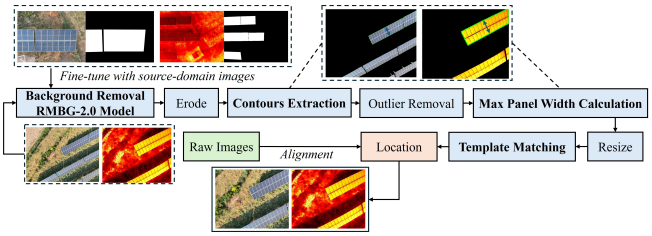

Figure 4: Spatial alignment for visible-infrared images.

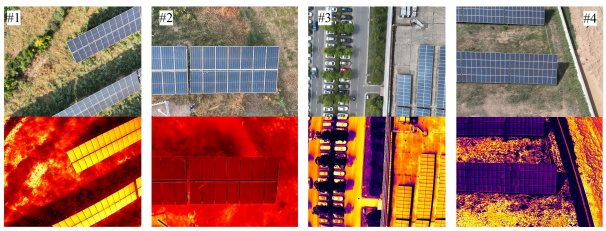

Figure 5: Spatial alignment results of #1~#4 domains.

of DD-LIVM, which consists of three main parts. We will first elaborate on our novel spatial alignment algorithm for infrared and visible images and how to preprocess them in § 3.1. Then, we will describe the model details and our defect-specific three-step fine-tuning (DTFT) strategy based on alternating modality masks and contrastive learning, in § 3.2, which serves as our core technical breakthrough to enable cross-domain detection. Moreover, we will present the source data augmentation and adaptive detection head selection schemes based on HSTD information in § 3.3, for further generaliability enhancement.

3.1 Spatial Alignment and Preprocessing

Before fine-tuning models to fuse infrared and visible feature maps, spatial alignment of the two modalities is essential. Due to the differences in the position, resolution and distortion of infrared and visible cameras, the practically collected two-modal images are misaligned. And this misalignment varies across different domains (*e.g.*, differing camera models, installation offsets, or drone altitudes). For the target domain with no prior information, it's hard to achieve effective alignment because 1) both feature-based and region-based alignment methods [13] struggle due to cross-modal heterogeneity in textures and intensity profiles; 2) deep learning-based pixel-level alignment modules in past works [43, 44] cannot generalize to new domains under varying view and position offsets.

To address this challenge, we propose a universal spatial alignment algorithm based on the consistency of PV panel widths on two-modal images, without knowing the internal and external parameters of camera or drone altitudes in new domains. Here, "consistency" denotes that while there is a scaling ratio gap between two-modal images collected at one time, the physical size of an individual PV panel is consistent across both modalities. This consistency applies to each panel individually and does not require different panels to have the same size.

The target of our algorithm is to get the *scaling and position offsets* between dual-modal images. As shown in Figure 4, our algorithm contains the following steps. First, we fine-tune the SOTA open-source model (RMBG-2.0 [52]) for **background removal** and retain only the PV panels' regions, based on the source-domain labels of PV panel's locations. Specifically, the RMBG-2.0 model is fine-tuned with the following loss:

$$\mathscr{L}_{br} = \alpha_1 \cdot \mathscr{L}_{SSIM}(I, I') + \alpha_2 \cdot \mathscr{L}_{BCE}(I, I') + \alpha_3 \cdot \mathscr{L}_{L1}(I, I') \quad (1)$$

where I is the raw image regardless of modalities, I' is the ground truth mask image where only PV panels' regions exist. $\mathscr{L}_{SSIM}$, $\mathscr{L}_{BCE}$ and $\mathscr{L}_{L1}$ denote the dimensionless structural similarity, binary cross entropy and absolute difference for pixels between I and I', respectively. The weights α_1, α_2 and α_3 follow the same setting in [52].

Accordingly, we can obtain the mask images of two modalities. Next, we erode the mask images to eliminate edge glitches for precise contour extraction. Then, we can derive the **contours of PV panels** in dual-modal images respectively, and possible outlier regions as well, which could be removed according to the region area. As we fine-tuned the background removal model to our PV scenario with sufficient images from source-domain sites, the contour extraction after background removal can achieve fine performance. Following this, we can get the minimum circumscribed rectangle of PV panels' contours in two images and find the ones with **maximum PV panel width**. By comparing the maximum widths of visible and infrared images, we can calculate the scaling value between dual-modal images. Then, after resizing the image with smaller scaling, we can conduct **template matching** for dual-modal images to match the locations of non-masked PV panels, and crop this area for alignment. In this work, the optical axes of the dual cameras onboard the drone are parallel. Our algorithm can also be easily extended to the cases of non-parallel optical axes by adding the rotation matrix estimation to reinforce template matching. Note that these operations are all conducted on mask images after background removal, instead of raw images. By saving the cropped areas' location and re-scaling it, we can extract the corresponding area on the raw images, thereby achieving the dual-modal spatial alignment in different domains.

Figure 5 demonstrates the spatial alignment results of four different domains. Compared with the raw images in Figure 2, we can see the effectiveness of our proposed spatial alignment algorithm, when the background is not complex and contains fewer objects. Correspondingly, the ground truth of PV defects will be adjusted based on the aligned region. The cropped part will not affect the defect detection because during the inspection, the drone moves slowly to the top of each PV panel to capture images, and each defect will appear in both modalities' images at a certain time without missing. Next, we conduct preprocessing operations on the aligned dual-modal images. We normalize the pixel value of infrared and visible images respectively and resize them to the unified shapes. Besides, we design augmentation methods on them (introduced in § 3.3), not limited to traditional ways like cutout or mixup [49].

3.2 Defect-specific Three-step Fine-tuning

The aligned and processed infrared-visible images will be inputted into our defect detection model.

Model Structure. As mentioned, our model exploits the infrared FM encoder from InfMAE [26] and the visible FM encoder from FM-FSOD [14] as backbones to extract feature maps. The infrared encoder has three blocks with output feature maps with different shapes: 112×122×256, 56×56×384 and 28×28×768. Then, the last-layer feature map can be downsampled to 14×14×768 by a convolution layer with a kernel size of 3, stride of 2 and padding of 1. Here applying multi-scale feature maps is necessary, because our 10 types of defects have different sizes. For example, the solar reflection defect has a large region but the dust defect has an extremely small size. Shallow and high-resolution feature maps can benefit detecting small-size defects while deep and low-resolution feature maps are more suitable to search large-size defects and extract defect semantics. Meanwhile, for visible FM encoder with 24 vision transformer (ViT) blocks, since the output of each ViT block is with the same resolution, we accordingly extract the 4, 8, 12 layers' output feature maps and add them to the 24-layer output, with weights of 0.5, 0.2 and 0.1 respectively. Thus, besides the final-layer output itself, we can obtain other three feature maps with the shape of 56×56×1024 and then downsample or upsample them to 112×112×1024, 28×28×1024 and 14×14×1024 through average pooling or unpooling layers, which corresponds to the infrared feature maps. Then, we are able to concatenate them to the dimensions of 1280, 1408, 1792 and 1792 during feature fusion and form four-scale infrared-visible feature maps for defect detection with different defect sizes. By above combining visible and infrared FM encoders and concatenating their feature maps, we obtain the basic structure of our large infrared-visible model. This model is dedicated to PV defect detection after fine-tuning, rather than a foundation model for various tasks.

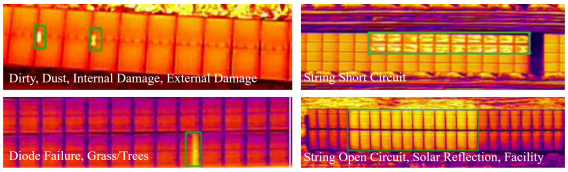

Figure 6: Defect-induced four hotspot shapes.

However, the challenge here is that, direct concatenating and fine-tuning the model does not perform well and yield overfitting to source-domain training data. Different from object detection in autonomous driving or remote sensing, *PV defects have special visible-infrared semantics.* For example, what infrared images capture is the hotspot, instead of the same physical object entity with visible images. As shown in Figure 1, the internal damage can cause the hotspot but is invisible (thus not appearing in visible images). So, intuitive fine-tuning will push the visible encoder to excessively learn the incorrect location contextual information, leading to source-domain overfitting.

To this end, we investigate the dual-modal semantics of PV defects [18, 41] and get two observations. On one hand, infrared images can capture hotspots with all defects' location information, but with only four low-resolution shapes (point/block, strip, bar and region). As shown in Figure 6,

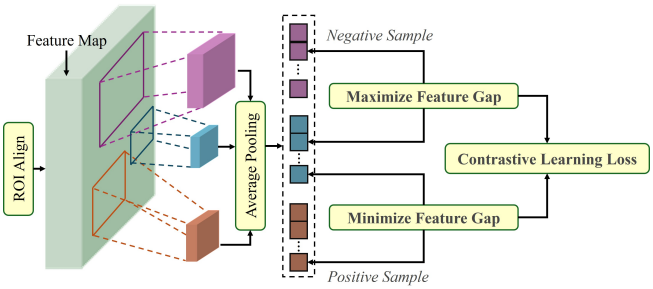

Figure 7: The working diagram of contrastive head.

dirty, dust, internal or external damage will induce point-like or small block-like tiny hotspots. Diode failure and grass/trees obstruction yield strip-like hotspots due to their inherent shapes. String short circuit makes bar-like larger hotspots with uneven temperature distribution. String open circuit, facility obstruction and solar reflection can cause a large area paralyzed, thus having region-like hotspots. So, infrared images are hard to distinguish the specific defect forms, for example, the dirty and internal damage. On the other hand, visible images contain only external defects' information including grass/trees, dirty, facilities, dust, external damage and solar reflection, while internal defects including diode failure, internal damage, string open circuit and short circuit are totally invisible. Therefore, the two modalities have separate semantics utility regarding the localization visibility and classification granularity of different partial defects. Based on this, we can add a modality masking mechanism to alternately fine-tune one of the infrared or visible encoders while masking the other. This envision can steer the optimization of dual-modality feature maps toward their respective semantic strengths, thereby helping reduce overfitting risks to the source domain.

Three-step Fine-tuning Strategy via Alternating Modality Masks. To transform the above idea into reality, first we incorporate a novel contrastive learning head for regulating feature gaps of different defects (see Figure 7) besides standard detection heads. According to the ground truth location of bounding box, we can obtain its corresponding features in the feature map through ROI align [16]. And based on the ground truth defect type, we can get positive and negative samples for each defect sample. We utilize the features' Euclidean distance after average pooling to the shape of $1\times1\times C$ to measure the feature similarity where C is the feature map's dimension, as the bounding boxes of different defects have diverse spatial sizes which are also probably not proportional. Then, we conduct the three-step fine-tuning as follows.

Step 1: Mask the visible encoder, fine-tune the infrared encoder. We regroup the 10 types of defects into 4 new categories according to their shapes in Figure 6 and construct positive-negative samples based on these 4 categories. Then, as the pink branch in Figure 3, we lead the infrared encoder

Algorithm 1: DTFT Strategy.

Input: Data: Dual-model training image set $\mathcal{I}_{ir}$ and $\mathcal{I}_{vis}$, corresponding ground truth bounding box set $\mathcal{P}$ and class label set C; Model: The infrared encoder $\Phi_{ir}(\cdot)$, the visible encoder $\Phi_{vis}(\cdot)$. Head losses: $\mathscr{L}_{ir}(\cdot)$, $\mathscr{L}_{vis}(\cdot)$ and $\mathscr{L}_{full}(\cdot)$ including detection head and contrastive head losses.

Output: Optimized encoders $\Phi'_{ir}(\cdot)$ and $\Phi'_{vis}(\cdot)$, and overall detection head.

```
Initialize pretrained infrared and visible encoders Φ_ir(·) and Φ_vis(·).
while iteration=1, 2, ..., I do
    if iteration ≤ γ then
        Update Φ_ir(·) by minimizing the loss of
          L_ir(Φ_ir(I_ir), P, C) with 4-class feature regulation based on hotspot shapes.
        Update Φ_vis(·) by minimizing the loss of
          L_vis(Φ_vis(I_vis), P, C) with 7-class feature regulation where internal defects' locations in P are used.
        Fuse the feature maps Φ_ir(I_ir) ⊕ Φ_vis(I_vis), and Update Φ_ir(·), Φ_ir(·) and overall detection head by minimizing the loss of
          L_full(Φ_ir(I_ir), Φ_vis(I_vis), P, C).
    else
        Freeze two encoders Φ_ir(·) and Φ_vis(·) and fine-tune the overall detection head.
    end
end
```

to the existing detection head and our contrastive head. The detection head oversees the hotspot regions localization and 4-class coarse-grained classification, with the detection loss $\mathscr{L}_{det}$. The contrastive learning head is used to enlarge the feature gap of distinct-shape defects and weaken that of same-shape defects. The overall loss can be formulated as:

$$\mathscr{L} = \beta_1 \cdot \mathscr{L}_{det} + \beta_2 \frac{1}{N} \sum_{i=1}^{N} \left\| F(P) - F(P_i^+) \right\|^2 - \beta_3^{ir} \frac{1}{M} \sum_{j=1}^{M} \left\| F(P) - F(P_j^-) \right\|^2 \quad (2)$$

where the $F(P)$ denotes the features at position P. $F(P^-)$ and $F(P^+)$ are features of negative samples and positive samples, where N and M denote their quantities in a batch respectively. The weight setting of β_1, β_2 and β_3 refers to § 4.1.2.

Here for the positive samples, we add a *soft constraint* instead of fully shortening their gaps, since same-shape defects may still have different sizes displayed in infrared images (*e.g.*, grass/trees). That is, for each defect, if the same-shape defect with a size difference larger than half of itself, we will not regard it as positive sample.

Step 2: Mask the infrared encoder, fine-tune the visible encoder. Similarly, we can regroup all the 4 internal defects into

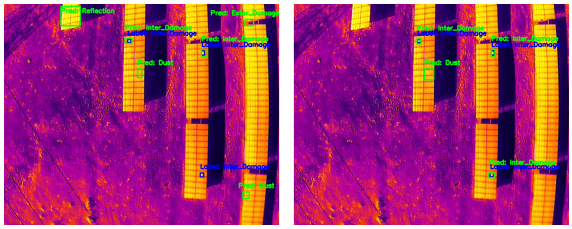

Figure 8: The cross-domain detection results w/o and w/ using DTFT strategy.

the same category and act as negative samples to the other 6 external defects. Following the blue branch in Figure 3, the detection head focuses on the regression of 6 external defects' locations and their classification. The regrouped category of internal defects is only used in contrastive head instead of classification task, and their features' positions are derived by the ground truth of aligned infrared images. Then, we can get a similar loss to Equation 2 with the consideration of feature gaps between mutual external defect categories, and between them and the internal defect category.

Step 3: Feature fusion and jointly fine-tune two encoders. After the first two steps' separate fine-tuning, we follow the black regular route in Figure 3, and conduct all defects' detection and feature regulation (step 3). During the training of step 3, the fused feature maps will be further optimized towards: 1) enhancing the localization information of external defects in dual-modal images; 2) fusing incomplete defect classification information of each modal in a fine-grained manner. This step is performed after semantic-level separate fine-tuning in step 1 and 2, and thus will not incur semantic mismatch or source-domain overfitting.

For every iteration in the fine-tuning process, we conduct the above three steps for constant semantic-level fusion of two modalities, as summarized in Algorithm 1. When calculating the feature distances in Equation 2, we will compute the average values on distinct scale's feature maps. To further prevent overfitting, After γ epochs, we would freeze the dual-modal encoders and only fine-tune the detection heads for the rest $K-\gamma$ epochs. The detection heads used are FCOS [42] and MaskRCNN [17], and we have an adaptive selection scheme for the two head types, proposed in § 3.3. Finally, we can obtain generalizable and defect-specific infrared-visible feature maps and enable cross-domain detection.

Figure 8 shows the cross-domain detection results using our three-step fine-tuning strategy with FCOS head. As compared to that of direct fine-tuning, we can find the decreases in false detection and missed detection samples. For example, the incorrect detection of solar reflection, probably caused by the visible encoder's source-domain overfitting, would not occur via our strategy. Nevertheless, there may still exist some incorrect cases such as external damage in Figure 8. This is potentially because the hotspot detection with tiny sizes is still vulnerable to variant domain shifts [32]. Therefore, in the next subsection, we will conduct data augmentation and head selection as assistance to enhance this.

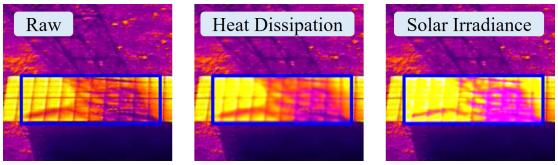

Figure 9: HSTD-based data augmentation samples.

3.3 HSTD-based Data Augmentation and Head Selection

While current infrared FM accounts for diverse background and contextual variations, industrial applications like PV defect detection face domain shifts far exceeding conventional vision scenarios due to intricate physical factors. For example, the solar irradiance intensity and ambient temperature critically influence PV module output power and heat dissipation efficiency, thereby altering hotspot morphology, intensity distribution, and spatial patterns [18, 41]. This makes the detection of tiny hotspots remain vulnerable to complex domain shifts.

To address this challenge, we propose to exploit the hotspot shape and temperature distribution (HSTD) information for source data augmentation and adaptive head selection, with the following details.

Source Data Augmentation. We simulate HSTD changes with environmental factors on infrared images' hotspots for data augmentation. First, considering the flight height related to image resolutions, we crop the images into 1/4 and 1/9 and then resample them back to the original resolution. For the solar irradiance, it affects the power consumption of defective panel. For the wind speed and ambient temperature, they affect the heat dissipation efficiency. Thus, we use the Laplace operator to simulate the process of heat diffusion and solar irradiance changes, and the pixel values of the hotspot area are updated through iterations. Specifically, we design a 3×3 Laplacian kernel ∇^2 for temperature diffusion to adjacent areas:

$$\frac{\partial T(x,y)}{\partial t} = \sigma \nabla^2 T(x,y),\ \nabla^2 = \begin{pmatrix} \eta_1 & \eta_2 & \eta_1 \\ \eta_2 & -1 & \eta_2 \\ \eta_1 & \eta_2 & \eta_1 \end{pmatrix} \tag{3}$$

where $T(x,y)$ denotes the temperature value at position (x,y) and $\sigma \in [0,1]$ is the thermal diffusivity to control the diffusion rate. $\eta_1, \eta_2 \in [0,1]$ reflect heat diffusion efficiency and σ reflects more about solar irradiance level. Figure 9 gives the examples of temperature diffusion under modified σ, η_1 and η_2. Thus, within five iterations' diffusion, we

can derive a series of augmented infrared images. Meanwhile, the augmented paired visible images are the same ones without changes. Traditional image transformations [49] including flip, rotation, Gaussian noise, and cutout are jointly used for data augmentation. Through this approach, the generalization of hotspot detection can be enhanced.

Adaptive Head Selection. Past domain alignment methods mostly follow the idea that models robust to perturbation, usually locate in the flat minima region of the parameter space and exhibit excellent generalization ability [51]. Inspired by visual in-context learning, we consider to add a same-type source-domain defect with minimum similarity to its nearby 50×50 pixels' region as the perturbation, for each detected defect in target-domain infrared images. If the detection result of this defect does not change after perturbation, we deem that this detection model is robust and suitable to the target domain. To measure the similarity and choose the perturbation sample, we design an HSTD-based distance between target sample and perturbation sample, from the perspectives of $\frac{S_{contour}}{W \cdot H}$ and $\frac{C_{contour}}{2W+2H}$ where $S_{contour}$ and $C_{contour}$ are the size and perimeter of one defect's contour after applying SAM model [22] to segment the defect' bounding box with width W and height H. The source-domain defect whose sum distance from the two perspectives is maximum, is chosen as the perturbation sample, and its region after SAM segmentation is added to the target image.

Based on this, we can select the model suitable for target domains. First, in § 3.2, We train our model with two types of heads (FCOS and MaskRCNN). As mentioned, after γ epochs, the dual-modal encoders will stop fine-tuning and be frozen. Then, in the latter head fine-tuning, we save the heads every 10 epochs, until K epochs. Accordingly, we can measure the robustness of $2 \cdot (K - \gamma)/10$ heads, through adding the perturbation. We calculate the average intersection over union (IoU) before and after perturbation, for all detected defects. Then we select the head with the maximum IoU. In this way, we can obtain the detection head with the least feature distribution discrepancy across domains.

4 Performance Evaluation

4.1 Experiment Setup

4.1.1 Datasets. Through drone-based inspection, we collect real-world infrared and visible images in four cities' practical PV power stations (see Figure 10) under totally nine different settings, denoted by #1∼#9. The dual-modal images' quantities in the nine scenarios are: 524, 296, 518, 2068, 1020, 842, 994, 566, 250 respectively, forming the dataset of totally 7078 dual-modal images. And the numbers of PV defects appearing in the images are 612, 563, 548, 426, 1627, 973, 1958, 192, 398 respectively (7297 defects in total with 10 defect types).

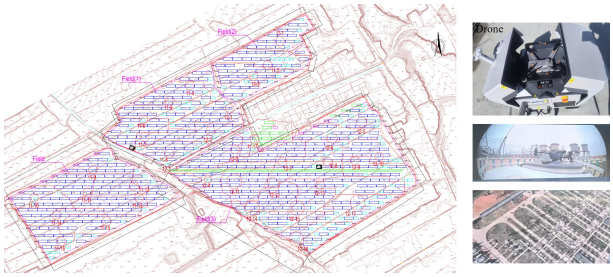

Figure 10: Topographic map of a practical PV station site for data collection, where the blue area denotes the PV panel's positions.

Table 1: Dataset Overview: for each target scenario, same and different domain factors with the other eight scenarios. "ST", "C", "P", "H", "T", "S" denote the factors of site terrain, camera device types, camera positions, drone flight heights, ambient temperature and solar irradiance, respectively."T"+ or "S"+ denote the higher ambient temperature or solar irradiance (specific values are not controllable). Besides, the image backgrounds vary across different scenarios.

Domain	Same Factors	Different Factors
#1	"ST" (#2)	"C", "P", "H", "T", "S"
#2	"ST" (#1)	"C", "P", "H", "T", "S"
#3	-	"ST", "C", "P", "H", "T", "S"
#4	"S"+	"ST", "C", "P", "H", "T"
#5	"S"+	"ST", "C", "P", "H", "T"
#6	"C", "ST" (#7), "T"+	' "P", "H","S"
#7	"C", "ST" (#6), "T"+	' "P", "H","S"
#8	"C", "ST" (#9)	"P", "H", "T", "S"
#9	"C", "ST" (#8)	"P", "H", "T", "S"

For #1∼#9 scenarios, a broad spectrum of factors are covered: site terrain, backgrounds, camera devices and positions, drone flight heights and views, ambient temperature and solar irradiance, with different domain shift degrees. Specifically, the data in #1∼#2 are collected in the same station with plain terrain, but with diverse camera positions, flight heights and sunshine conditions (morning and afternoon, respectively). #3 is independently collected over the rooftop at noon. As compared to #1∼#3, besides the basic gaps of separate sites with varying camera models, backgrounds and climates, the individual differences of #4∼#9 are as follows. #4 and #5 are under much stronger solar irradiance, #6 and #7 are under higher ambient temperature, and #8 and #9 are in the sandland terrain. The discrepancy between #4 and #5 lies in camera positions, flight heights and climate conditions, which also applies to #6 and #7 as well as #8 and #9. Note that the real-world PV defects need to be repaired as soon as possible, and thus we are unable to control the variable of individual environmental factor during data collection. For

each target scenario, the same and different domain factors with the other eight scenarios are summarized in Table 1.

4.1.2 Model Implementation and Training. For the spatial alignment module, we fine-tune the RMBG-2.0 model [52] for background removal, using the AdamW optimizer with a learning rate of 10^{-6} and batch size of 16. The loss weights α_1, α_2 and α_3 are set to 0.1, 1 and 0.3, respectively. We train it on an NVIDIA A100 GPU for 100 epochs. For the defect detection model, the infrared backbone (*i.e.*, InfMAE encoder [26]) is pretrained on Inf30 dataset, while the visible backbone (*i.e.*, FM-FSOD encoder [14]) is pretrained on LVD-142M dataset. We load the two pretrained encoders to our DD-LIVM. Then, we use the neck of FPN [25] and detection heads (FCOS [42] and MaskRCNN [17]) in MMdetection Toolbox [6] for defect detection. The neck's input dimensions are modified to 1280, 1408, 1792 and 1792 to to be consistent with our fused feature maps, while its output remains 256 dimensions for the heads. Our DFTF strategy is conducted for every batch of data. For the first two steps' fine-tuning, we similarly change the necks' input dimensions for the infrared branch and the visible branch based on their separate feature map shapes, respectively. During the overall fine-tuning, we use an AdamW optimizer with an initial learning rate of 0.0001 and weight decay of 0.1. The dimensionless loss weights are $\beta_1 = 0.8$, $\beta_2 = 0.2$, $\beta_3 = 0.2$, and the data augmentation parameters are $\eta_1 = 0.1$, $\eta_2 = 0.4$ with $\sigma \in [0.1, 0.3]$ randomly generated during augmentation. Then, we set the batch size to 4 and train our detection model for 300 epochs for convergence. After 200 epochs, we freeze the two encoders' parameters, and only fine-tune the neck and detection head for the extra 100 epochs. Thus, we can get 10 FCOS heads and 10 MaskRCNN heads for selection. The detection models are trained on 8× NVIDIA A100 GPU.

4.1.3 Baselines. We compare the performance of DD-LIVM with the following baselines.

Two SOTA defect detection methods: (1) NIF [19] uses YOLOv5 to segment the visible PV array images into PV modules and get the corresponding infrared ones. Then, it exploits ResNet to detect the defects on PV modules' infrared images for coarse-grained diagnosis. (2) CDPC [5] employs generative adversarial networks to augment the collected dual-modal images and conducts data selection with pseudo-label cross-entropy for CNN detection model training.

Three SOTA infrared-visible object detection methods: (1) DAMSDet [12] employs a multispectral deformable cross-attention module to adaptively sample and aggregate multi-semantic level features for each object, thus realizing promising object detection. (2) GM-DETR [47] utilizes self-attention operations on the dual-modal top-level CNN features and then effective feature fusion across multiple scales, thereby greatly promoting the object detection performance. (3) DPDETR [11] proposes a decoupled position multispectral cross-attention module for complementary feature fusion with the constraint of dual-modal reference positions, thus achieving more robust object detection under spatial misalignment.

The above 5 baselines are identically inputted with our infrared-visible image pairs and output the detection results of defects for model training or inference.

4.1.4 Evaluation Metrics. Referring to past defect detection works, we exploit the metrics: (1) **detection accuracy** [5, 19] and (2) **mean average precision (mAP)** with 0.4 IoU threshold (mAP@40) [9, 21] for performance evaluation. For the detection accuracy, it means that for each predicted defect's bounding box, we find the one with the maximum IoU value in all ground truth defect bounding boxes in the image, and then compare their defect classification results among 10 types. If the classification result is correct, we regard it as a correct sample. For those predictions with zero IoU or false classification results, we regard them as false samples. For the mAP@40, this chosen IoU threshold of 0.4 is consistent with existing works. Since the end goal of PV defect detection is to let workers go and repair these defects on-site, fine-grained localization precision with a very large IoU threshold (*e.g.*, mAP@0.9) is not necessary in practical applications. Instead, generally an IoU threshold of 0.4 or 0.5 is enough for PV maintenance demands [9, 21].

4.2 Overall Performance

For each scenario, we use the other 8 scenarios' images for training (source domain) and this scenario's images for testing (target domain) for cross-domain evaluation. The data in the target domain is completely unseen samples, with diverse domain shifts from the source domain.

4.2.1 Cross-domain Performance with Baseline Comparison. The overall prediction performance is shown in Table 2 for all nine scenarios. As we can see, our DD-LIVM model achieves an average detection accuracy of 87.7% across nine scenarios, greatly surpassing the SOTA schemes by 17.3%. As for the mAP@0.4, DD-LIVM can reach 80.6% while the SOTA scheme is only 64.7%. This is reasonable because the feature maps of DD-LIVM fit the defect semantic better and involve more domain-invariant information to avoid source-domain overfitting. In contrast, GM-DETR [47] and DPDETR [11] are designed with the complementary paradigm of infrared and visible features, thus not compatible with defect semantics. Besides, the DD-LIVM's accuracies under all nine scenarios are over 80%, indicating its fine generalizability. Among them, #2, #8 and #9 scenarios perform better than other scenarios, probably due to more training data increasing source diversity. Differently, #3 scenario's performance is limited

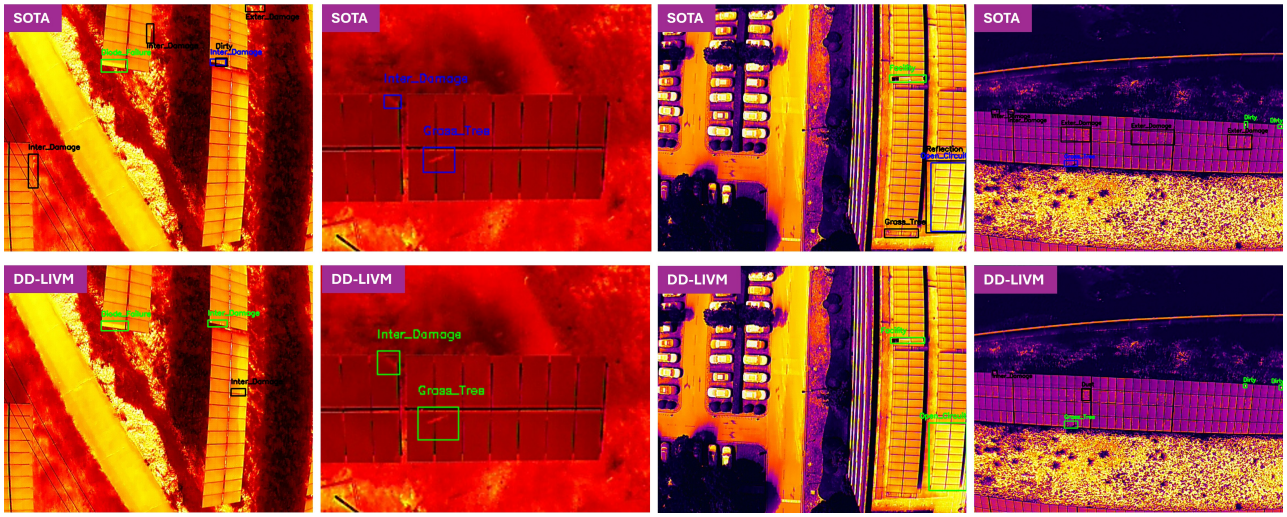

Figure 11: Cross-domain detection visualization for #1~#4 scenarios. Green means the correct detection samples, Blue means false negative samples and Black means false positive samples.

Table 2: Comparison with (1) SOTA defect detection methods and (2) SOTA infrared-visible object detection methods under #1~#9 scenarios in our datasets. The detection accuracy is given, with mAP@0.4 in the bracket as well.

Model	#1	#2	#3	#4	#5	#6	#7	#8	#9
NIF [19]	0.27(0.19)	0.22(0.18)	0.33(0.25)	0.35(0.27)	0.38(0.29)	0.25(0.16)	0.21(0.18)	0.42(0.36)	0.43(0.32)
CDPC [5]	0.52(0.41)	0.41(0.37)	0.42(0.36)	0.38(0.31)	0.36(0.32)	0.52(0.43)	0.55(0.53)	0.55(0.42)	0.59(0.47)
DAMSDet [12]	0.68(0.49)	0.62(0.59)	0.54(0.48)	0.50(0.43)	0.51(0.43)	0.47(0.32)	0.53(0.50)	0.65(0.58)	0.63(0.57)
DPDETR [11]	0.71(0.61)	0.75(0.71)	**0.63(0.60)**	**0.64(0.53)**	0.60(0.47)	0.56(0.42)	0.60(0.57)	0.74(0.64)	**0.79(0.71)**
GM-DETR [47]	**0.74(0.68)**	**0.76(0.74)**	0.60(0.58)	0.62(0.50)	**0.70(0.62)**	**0.72(0.71)**	**0.70(0.68)**	**0.76(0.64)**	0.75(0.68)
DD-LIVM	**0.87(0.80)**	**0.94(0.91)**	**0.83(0.77)**	**0.82(0.74)**	**0.84(0.75)**	**0.90(0.82)**	**0.85(0.84)**	**0.92(0.80)**	**0.93(0.83)**

Table 3: Comparison of Accuracy and Model Overhead

Model	Accuracy	Params (M)	FLOPs (G)
DPDETR [11]	0.669	90.1	208.0
GM-DETR [47]	0.704	70.3	175.8
DD-LIVM (Alignment)	/	**220.2**	**533.6**
DD-LIVM (Detection)	**0.877**	**419.5**	**963.2**

by terrain background gaps from plain to rooftop, which induces large domain shifts. #4 and #5 also suffer from a relatively low accuracy possibly due to its increased solar irradiance levels beyond the model's ability to resolve hot spots. Despite that, the performance gains of #3, #4, #5 scenarios compared to baselines are significantly evident with 14%~23% accuracy improvement, which demonstrates our designs' superiority for cross-domain detection.

Moreover, from Figure 11, we can clearly observe that the SOTA scheme leads to a lot of false detection and missed detection phenomenons. For example, both internal and external defects are missed in #2 while excessive false detections of external damages occur in #4, primarily due to the ineffective dual-modal fusion on defect locations' visibility - rooted in the unique semantics of PV scenarios. Instead, DD-LIVM successfully addresses this through PV-specific DTFT strategy based on alternating modality mask, thus showing satisfactory detection results. Additionally, we find that some false positive samples occur at the edge of a PV panel that may be confused with edge damage or grass/trees obstruction samples, or affected by imprecise spatial alignment. We may adopt a two-stage detection strategy to address this, that is, firstly locating PV panels, segmenting them, and then detecting the defects on each panel. This can avoid incorrect detection at edges and also enhances tiny defect detection. Nevertheless, this work does not follow this and employs the idea of multi-scale feature maps alternatively, as this operation may lose contextual information such as background temperatures or accumulate the errors from PV panel localization. We will explore its trade-off to further promote our detection precision in our future works.

4.2.2 Performance of Different Defect Types. As given in Table 4, we count each defect type's accuracy. It is demonstrated that different types of PV defects have distinct detection performances. As compared to DPDETR [11], both

Table 4: Performance of Different Defect Types

Type	DPDETR [11]	GM-DETR [47]	**DD-LIVM (Ours)**
Grass/Tree	0.783	0.818	**0.861**
Facility	0.841	0.823	**0.937**
Dirty	0.458	0.669	**0.844**
Dust	0.503	0.634	**0.763**
Exter Damage	0.626	0.691	**0.872**
Solar Reflection	0.813	0.868	**0.956**
Short Circuit	0.741	0.754	**0.913**
Open Circuit	0.656	0.605	**0.925**
Inter Damage	0.547	0.490	**0.775**
Diode Failure	0.728	0.742	**0.933**

GM-DETR [47] and our DD-LIVM model focuses more on multi-scale feature maps fusion, thus enhancing the detection precision of tiny objects such as dust, dirty and inter damages, which are always hard to distinguish across various domains, especially with temperature (color depth of infrared images) or drone's height (image resolution) changes. The grass/tree, facility obstruction and solar reflection are manifest, thereby holding a high accuracy of over 80% even for baselines. Due to little consideration of GM-DETR in spatial offsets across domains, it leads to a series of incorrect detection of inter damages and extra damages, where the model overfits to the incorrect visible features. By contrast, DD-LIVM markedly promotes the accuracy of all defect types through semantic-level fine-tuning and fusion, especially for internal four defects. This is because our DTFT strategy can help differentiate the external defects' features apart from those similar ones of internal defects in the infrared images, by conducting separate fine-tuning before feature fusion.

4.2.3 Model Size and Overhead. As shown in Table 3, for spatial alignment, DD-LIVM exploits the fine-tuned RMBG-2.0 model based on BiRefNet [52] for removal background. This model has a total parameter number of 220.2M with 533.6G floating-point operations per second (FLOPs). For the subsequent defect detection, DD-LIVM merges the infrared and visible FM encoders, thereby inducing 419.5M parameter quantity with 963.2G FLOPs. The overall inference time of DD-LIVM for an infrared-visible image pair on an NVIDIA A100 GPU is 853.47ms, which meets the requirement of low-latency usage. Since our models are deployed offline and used after the drone inspection collects all the images, the increased storage space and larger FLOPs of DD-LIVM are acceptable on the server side. Note that, in practical PV sites, currently the defect detection is all conducted after drone inspection is completed and all collected images are uploaded to the server. We follow this way and it does not need edge deployment on the drone.

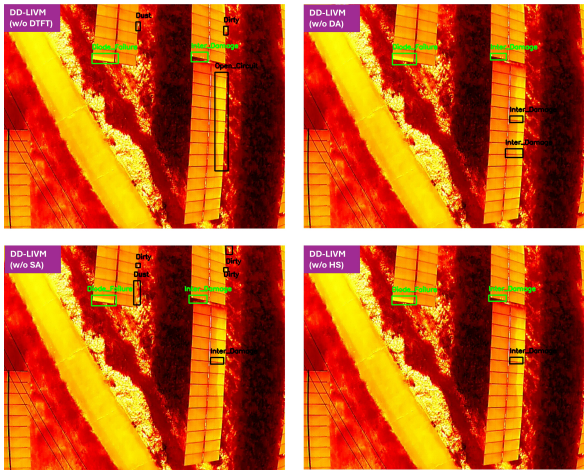

Figure 12: Ablation study visualization for #1 scenario, with comparison to Figure 11.

4.3 Ablation Study

Then we conduct an ablation study to verify the effectiveness of each module of system designs in DD-LIVM (see Table 5).

4.3.1 Effectiveness of Our DTFT strategy. As compared to the results of intuitive fine-tuning on two FM encoders, DD-LIVM employing DTFT can promote the average detection accuracy by 10.2%, indicating the effectiveness of semantic-level fine-tuning. For example, as depicted in Figure 12, our model without DTFT may excessively rely on the incorrect contextual information of string open circuit in the visible images during training, yielding false positive samples during inference. By using DTFT, the understanding of features' semantics will follow its physical mechanisms and circumvent overfitting. This conclusion holds for all nine scenarios, with a few differences in accuracy gains due to distinct defects' distributions. In the scenario containing more defects with difficulty to tell apart from only single-modal images, the accuracy improvement would be greater. Moreover, even without DTFT, DD-LIVM still outperforms the SOTA scheme GM-DETR by 7.1%, due to the usage of FM encoders as backbones and inherent advantages of FMs' generalizability.

4.3.2 Effectiveness of Our Spatial Alignment Algorithm. If deleting our spatial alignment module and employing the SOTA pixel alignment model [10] for this task, the performance will be apparently degraded by 9.4%. This is because DD-LIVM strongly hinges on the spatial alignment for its contrastive head during DTFT. The feature mismatching will greatly affect the regulated feature distances in contrastive learning, resulting in a bad fusion of two modalities. Besides, the misaligned external defects' features will lead to model overfitting. Yet, past pixel alignment models cannot generalize to new domains with totally different camera views. Instead, our alignment algorithm, based on the consistency of

Table 5: Ablation study for four modules of DD-LIVM: Three-step fine-tuning strategy (DTFT), Spatial alignment algorithm (SA) and HSTD-based data augmentation (DA) and head selection (HS).

Model	#1	#2	#3	#4	#5	#6	#7	#8	#9	Avg
GM-DETR (SOTA)	0.735	0.760	0.599	0.620	0.697	0.718	0.702	0.758	0.754	0.704
DD-LIVM (w/o DTFT)	0.801	0.832	0.709	0.689	0.735	0.819	0.757	0.806	0.824	0.775
DD-LIVM (SA→[10])	0.795	0.826	0.751	0.721	0.794	0.757	0.714	0.825	0.868	0.783
DD-LIVM (DA→OA-DG [23])	0.849	0.891	0.818	0.763	0.757	0.880	0.846	0.854	0.876	0.837
DD-LIVM (HS→MaskRCNN)	0.817	0.909	0.776	0.758	0.745	0.850	0.834	0.822	0.841	0.817
DD-LIVM (HS→FCOS)	0.868	0.935	0.796	0.781	0.828	0.883	0.852	0.877	0.885	0.856
DD-LIVM	**0.868**	**0.935**	**0.832**	**0.817**	**0.836**	**0.904**	**0.852**	**0.920**	**0.928**	**0.877**

physical PV panel widths, can apply to diverse new domains with strong interpretability and generalizability. Figure 12 clearly shows how our alignment module benefits reducing the false detection cases at PV panels' edges.

4.3.3 Effectiveness of Our HSTD-based Data Augmentation. As compared to the SOTA data augmentation method for object detection, *i.e.*, OA-DG [23], our data augmentation based on HSTD owns a finer accuracy of 4%, as the small hotspots which need augmentation most are vulnerable to domain shifts of solar radiation and ambient temperature changes. These augmentation factors are related to thermal infrared characteristics, which past works do not consider. For example, OA-DG cares only the foreground and background instances. Thus, our HSTD-based data augmentation is a more appropriate choice for our PV scenarios, which can also be combined with past methods for joint usage.

4.3.4 Effectiveness of Our HSTD-based Head Selection. DD-LIVM have 20 detection heads with two types (FCOS and MaskRCNN) trained for selection in different target domains. If removing the head selection module and directly using the FCOS head or MaskRCNN head at the last training epoch, the detection accuracy of DD-LIVM will be decreased by 2.1% and 6.0%, given that different source data with various distributions and quantities should have a different optimal training parameters including the number of epochs. So, through our head selection, we can further mitigate the source overfitting risks. Besides, in some domains (*e.g.*, #1, #2 and #7), DD-LIVM is not easy to overfit after semantic alignment and data augmentation, where our head selection scheme serves more as an insurance function. While for other domains like #8 and #9, the gains are evident. Additionally, we can see that under all nine scenarios, the FCOS heads all perform better than MaskRCNN heads, probably because the anchor settings in MaskRCNN are default hyperparameters that are not generic across different domains.

Overall, the DTFT strategy and spatial alignment are two essential modules for DD-LIVM's performance, which corresponds to our key contributions as well.

5 Related Work

In this section, we briefly review existing works from the following four perspectives.

PV Defect Detection. Based on the drone-captured dual-modal images, deep learning approaches are applied for detect detection [3, 20, 28, 38, 39, 50], while most of them directly utilize the universal models in the field of object detection. For example, PA-YOLO [50] uses YOLOv7 as backbone and adds an asymptotic feature pyramid network for feature fusion. Similarly, YOLOv5 and ResNet are used in [5]. These works do not consider two modalities' defect semantics in model design and only train their simple models on small-scale data from only a single site, thereby yielding very low generalizability to environmental condition changes. In contrast, DD-LIVM utilizes the latest FM encoders and employs three-stage fine-tuning on them according to defect semantics, thus greatly promoting the generalization performance.

Infrared-Visible Object Detection. As a similar task, current infrared-visible model frameworks for object detection [40] such as the SOTA ones GM-DETR [47] and DPDETR [11], adopt a complementary fusion paradigm for dual-modal features. That is, visible images provide the color and texture details while infrared images enhance low-light visibility of the same physical object. In these works, halfway fusion is always operated to merge the dual-modal complementary information. Thus, they perform poorly in PV defect scenarios and lead to model overfitting, since the dual-modal defect semantics are asymmetrical and not a simply complementary relationship.

Domain Generalization. For the prevailing domain generalization methods for object detection (*e.g.*, OA-DG [23], DivAlign [7]), whether through data augmentation, test-time adaptation or meta-learning strategies [53], their efficacy critically relies on exhaustive source-domain data diversity to achieve generalizable feature representations, since the feature map generalizability scales with the entropy of source-domain distributions [29]. They inherently require sufficient source data coverage to disentangle domain-invariant patterns from spurious correlations. However, this data diversity

remains constrained in PV defect datasets due to the difficulty of large-scale data collection in practical stations and annotation costs. Thus, the source diversity cannot cover various domain shifts in PV scenarios, such as frequent changes in solar irradiance or ambient temperatures, rendering existing domain generalization gains marginal in PV scenarios.

Foundation Models. Recent years have witnessed the rapid emergence of FMs in computer vision, which leverage large-scale pretraining to extract generalized features applicable to diverse downstream tasks [2, 14, 30, 46]. The feature extractors of FMs, *i.e.*, encoder modules, have been used as backbones for object detection frameworks, such as FOMO [54] and SyncOOD [27]. Currently, due to the lack of aligned visible-infrared image pairs and unaffordable overhead to collect for training, there is no infrared-visible multimodal FM. Thus, in this paper, we attempt to fuse the infrared FM encoder and visible FM encoder. And through our three-step fine-tuning based on defect semantics, we can obtain a large infrared-visible model dedicated to PV defect detection.

6 Discussion and Future Works

In this section, we will discuss the DD-LIVM's limitations and potential extensions.

New Defect Types. The design of DD-LIVM is not limited to this paper's ten common defect types, where our core innovation (*i.e.*, DTFT strategy) is scalable regardless of the defect quantity. When facing a new defect type, the defect regrouping criteria in DTFT could be accordingly updated based on the new defect's visibility in RGB images and its induced hotspot shapes and sizes in infrared images. Given this, integrating new defect types would only necessitate adjustments to the loss function computation, leaving the overall framework and training pipeline of DD-LIVM intact.

Defect Detection Precision. The effectiveness of DD-LIVM may still be affected by 1) extremely small defects under elevated flight altitudes; 2) sophisticated backgrounds or occlusion in new domains, where spatial alignment performance inevitably declines. Nevertheless, we can 1) conduct a two-stage detection by locating the PV panel first and detecting the defects on the segmented panel to enhance tiny object detection; 2) harness more diverse background images from other open datasets containing similar-size objects to PV panels to augment the fine-tuning performance of our background removal model in spatial alignment, preventing extracting the contours of other background objects. Besides, for occlusion issues, as the spatial alignment is actually for dual-modal cameras' calibration where the dual-modal scaling ratio and position offset are fixed in one domain, in all images during drone inspection, there must exist a few clean, unobstructed images. So, we can use these clean ones to compute the alignment parameters and use them for all images in one domain. These approaches will be explored in our future work. Actually, such noisy-image cases are not common, as most PV power stations are located in open areas.

Drone Operation Optimization. Currently, in practical PV sites, the drone is employed to collect the images at a fixed height along a pre-defined trajectory. We also follow this way in our work. As we can see, due to the size differences of diverse defect types, some tiny defects like dust and inter damages need finer-grained resolutions to locate and classify. Thus, in our future work, we will further exploit vision-language-action models to optimize the drone operation for data collection, such as decreasing the flight height when detecting the possible tiny defects, by the drone's autonomous decision-making.

Extension Applications. While DD-LIVM is designed for PV defect detection, the idea of DTFT strategy can be similarly applied to other infrared-visible inspection applications such as substation or mechanical equipment monitoring, where the two modalities process asymmetric semantics as well. Our two fine-tuned encoders could be further adapted to these similar applications as initial backbones. Besides, the insight of alternating modality mask behind DTFT can benefit the semantic alignment of the other image-wise modalities such as the depth images.

7 Conclusion

This paper introduces DD-LIVM, the first large infrared-visible model to enable cross-domain PV defect detection with infrared and visible images, without any prior information of new domains. To achieve this, DD-LIVM leverages infrared and visible FM encoders as backbones and proposes a three-step fine-tuning strategy based on alternating modality masks, thereby extracting generalizable infrared-visible feature maps that fit the specific semantics of PV defects. For practical employment of DD-LIVM, we further propose a universal spatial alignment algorithm for dual-modal images, and develop source data augmentation and adaptive detection head selection schemes based on infrared hotspot characteristics. We collect 7078 dual-modal images with 7297 defects in four cities' real-world PV power stations under nine different settings for evaluation. The experimental results demonstrate that DD-LIVM can achieve a high accuracy of 87.7% in cross-domain defect detection, outperforming the SOTA schemes by 17.3%. DD-LIVM delivers valuable insights into multimodal alignment and semantic-level fusion, and the utilization of foundation models in mobile systems.

Acknowledgments

This research is supported in part by RGC under Contract CERG 16206122, 16204523, 16205824, AoE/E-601/22-R, R6021-20, SRFS2425-6S05 and Contract R8015, and National Key R&D Program Project (2024YFE0203700).

References

[1] M Waqar Akram, Guiqiang Li, Yi Jin, Xiao Chen, Changan Zhu, Xudong Zhao, M Aleem, and Ashfaq Ahmad. 2019. Improved outdoor thermography and processing of infrared images for defect detection in PV modules. *Solar Energy* 190 (2019), 549–560.

[2] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2025. Foundation Models Defining a New Era in Vision: a Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

[3] Yacine Boutana, Sofiane Haddad, and Ammar Soukkou. 2024. Anomaly Detection and Classification of Solar Photovoltaic Modules Using Vision Transformers (ViT). In *International Conference on Artificial Intelligence in Renewable Energetic Systems*. Springer, 304–311.

[4] Gyusam Chang, Jiwon Lee, Donghyun Kim, Jinkyu Kim, Dongwook Lee, Daehyun Ji, Sujin Jang, and Sangpil Kim. 2024. Unified domain generalization and adaptation for multi-view 3d object detection. *Advances in Neural Information Processing Systems* 37 (2024), 58498–58524.

[5] XUEWEI CHAO, LIXIN ZHANG, YANG LI, CHAO HUANG, and JING LI. 2024. Photovoltaic fault detection based on infrared and visible image augmentation and fusion. *Turkish Journal of Agriculture and Forestry* 48, 3 (2024), 430–442.

[6] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. 2019. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155* (2019).

[7] Muhammad Sohail Danish, Muhammad Haris Khan, Muhammad Akhtar Munir, M Saquib Sarfraz, and Mohsen Ali. 2024. Improving Single Domain-Generalized Object Detection: A Focus on Diversification and Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17732–17742.

[8] DD-LIVM. 2025. DD-LIVM dataset samples of infrared-visible images. https://drive.google.com/drive/folders/1-EMVcw5exQqCqP5eeL8RqvrUreGYwHWE?usp=sharing.

[9] Antonio Di Tommaso, Alessandro Betti, Giacomo Fontanelli, and Benedetto Michelozzi. 2022. A multi-stage model based on YOLOv3 for defect detection in PV panels based on IR and visible imaging by unmanned aerial vehicle. *Renewable energy* 193 (2022), 941–962.

[10] Zhinan Gao, Dongdong Li, Yangliu Kuai, Rui Chen, and Gongjian Wen. 2025. Visible-Infrared Image Alignment For Unmanned Aerial Vehicles: Benchmark and New Baseline. *IEEE Transactions on Geoscience and Remote Sensing* (2025).

[11] Junjie Guo, Chenqiang Gao, Fangcen Liu, and Deyu Meng. 2024. DPDETR: Decoupled position detection transformer for infrared-visible object detection. *arXiv preprint arXiv:2408.06123* (2024).

[12] Junjie Guo, Chenqiang Gao, Fangcen Liu, Deyu Meng, and Xinbo Gao. 2024. Damsdet: Dynamic adaptive multispectral detection transformer with competitive query selection and adaptive feature fusion. In *European Conference on Computer Vision*. Springer, 464–481.

[13] Surbhi Gupta, Munish Kumar, and Anupam Garg. 2019. Improved object recognition results using SIFT and ORB feature detector. *Multimedia Tools and Applications* 78, 23 (2019), 34157–34171.

[14] Guangxing Han and Ser-Nam Lim. 2024. Few-shot object detection with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 28608–28618.

[15] Boyong He, Yuxiang Ji, Qianwen Ye, Zhuoyue Tan, and Liaoni Wu. 2025. Generalized Diffusion Detector: Mining Robust Features from Diffusion Models for Domain-Generalized Detection. *arXiv preprint arXiv:2503.02101* (2025).

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[18] Ula Hijjawi, Subhash Lakshminarayana, Tianhua Xu, Gian Piero Malfense Fierro, and Mostafizur Rahman. 2023. A review of automated solar photovoltaic defect detection systems: Approaches, challenges, and future orientations. *Solar Energy* 266 (2023), 112186.

[19] Feng Hong, Jie Song, Hang Meng, Rui Wang, Fang Fang, and Guangming Zhang. 2022. A novel framework on intelligent detection for module defects of PV plant combining the visible and infrared images. *Solar Energy* 236 (2022), 406–416.

[20] Muhammad Hussain and Rahima Khanam. 2024. In-depth review of yolov1 to yolov10 variants for enhanced photovoltaic defect detection. In *Solar*, Vol. 4. MDPI, 351–386.

[21] Rahima Khanam, Tahreem Asghar, and Muhammad Hussain. 2025. Comparative Performance Evaluation of YOLOv5, YOLOv8, and YOLOv11 for Solar Panel Defect Detection. In *Solar*, Vol. 5. MDPI, 6.

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

[23] Wooju Lee, Dasol Hong, Hyungtae Lim, and Hyun Myung. 2024. Object-aware domain generalization for object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 2947–2955.

[24] Hulin Li. 2024. Rethinking Features-Fused-Pyramid-Neck for Object Detection. In *European Conference on Computer Vision*. Springer, 74–90.

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

[26] Fangcen Liu, Chenqiang Gao, Yaming Zhang, Junjie Guo, Jinghao Wang, and Deyu Meng. 2024. InfMAE: A foundation model in the infrared modality. In *European Conference on Computer Vision*. Springer, 420–437.

[27] Jiahui Liu, Xin Wen, Shizhen Zhao, Yingxian Chen, and Xiaojuan Qi. 2024. Can OOD Object Detectors Learn from Foundation Models?. In *European Conference on Computer Vision*. Springer, 213–231.

[28] Qing Liu, Min Liu, Chenze Wang, and QM Jonathan Wu. 2024. An efficient CNN-based detector for photovoltaic module cells defect detection in electroluminescence images. *Solar Energy* 267 (2024), 112245.

[29] Shaocong Long, Qianyu Zhou, Chenhao Ying, Lizhuang Ma, and Yuan Luo. 2024. Rethinking domain generalization: Discriminability and generalizability. *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

[30] Siqi Lu, Junlin Guo, James R Zimmer-Dauphinee, Jordan M Nieusma, Xiao Wang, Steven A Wernke, Yuankai Huo, et al. 2025. Vision Foundation Models in Remote Sensing: A survey. *IEEE Geoscience and Remote Sensing Magazine* (2025).

[31] Nehemiah Mukwevho, Andile Mkhohlakali, Napo Ntsasa, James Sehata, Luke Chimuka, James Tshilongo, and Mokgehle R Letsoalo. 2025. Methodological approaches for resource recovery from end-of-life panels of different generations of photovoltaic technologies–A review. *Renewable and Sustainable Energy Reviews* 207 (2025), 114980.

[32] Muhammed Muzammul and Xi Li. 2021. A survey on deep domain adaptation and tiny object detection challenges, techniques and datasets. *arXiv preprint arXiv:2107.07927* (2021).

[33] KN Nwaigwe, Philemon Mutabilwa, and Edward Dintwa. 2019. An overview of solar power (PV systems) integration into electricity grids.

Materials Science for Energy Technologies 2, 3 (2019), 629–633.

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[35] Emir Ozturk, Emanuele Ogliari, Maciej Sakwa, Alberto Dolara, Nicola Blasuttigh, and Alessandro Massi Pavan. 2024. Photovoltaic modules fault detection, power output, and parameter estimation: A deep learning approach based on electroluminescence images. *Energy Conversion and Management* 319 (2024), 118866.

[36] Constantine P Papageorgiou, Michael Oren, and Tomaso Poggio. 1998. A general framework for object detection. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*. IEEE, 555–562.

[37] Bhubaneswari Parida, Selvarasan Iniyan, and Ranko Goic. 2011. A review of solar photovoltaic technologies. *Renewable and sustainable energy reviews* 15, 3 (2011), 1625–1636.

[38] EA Ramadan, Nada M Moawad, Belal A Abouzalm, Ali A Sakr, Wessam F Abouzaid, and Ghada M El-Banby. 2024. An innovative transformer neural network for fault detection and classification for photovoltaic modules. *Energy Conversion and Management* 314 (2024), 118718.

[39] Binyi Su, Haiyong Chen, and Zhong Zhou. 2021. BAF-detector: An efficient CNN-based detector for photovoltaic cell defect detection. *IEEE Transactions on Industrial Electronics* 69, 3 (2021), 3161–3171.

[40] Yuxuan Sun, Yuanqin Meng, Qingbo Wang, Minghua Tang, Tao Shen, and Qingwang Wang. 2023. Visible and infrared image fusion for object detection: a survey. In *International Conference on Image, Vision and Intelligent Systems*. Springer, 236–248.

[41] Wuqin Tang, Qiang Yang, Zhou Dai, and Wenjun Yan. 2024. Module defect detection and diagnosis for intelligent maintenance of solar photovoltaic plants: Techniques, systems and perspectives. *Energy* (2024), 131222.

[42] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9627–9636.

[43] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. 2019. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3623–3632.

[44] Guan-An Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zeng-Guang Hou. 2020. Cross-modality paired-images generation for RGB-infrared person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 12144–12151.

[45] Fan Wu, Jinling Gao, Lanqing Hong, Xinbing Wang, Chenghu Zhou, and Nanyang Ye. 2024. G-nas: Generalizable neural architecture search for single domain generalization object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 5958–5966.

[46] Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. 2024. General object foundation model for images and videos at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3783–3795.

[47] Yiming Xiao, Fanman Meng, Qingbo Wu, Linfeng Xu, Mingzhou He, and Hongliang Li. 2024. Gm-detr: Generalized muiltispectral detection transformer with efficient fusion encoder for visible-infrared detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5541–5549.

[48] Hanfei Xie, Baoxi Yuan, Chengyu Hu, Yujie Gao, Feng Wang, Chunlan Wang, Yuqian Wang, and Peng Chu. 2024. ST-YOLO: A defect detection method for photovoltaic modules based on infrared thermal imaging and machine vision technology. *PloS one* 19, 12 (2024), e0310742.

[49] Mingle Xu, Sook Yoon, Alvaro Fuentes, and Dong Sun Park. 2023. A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition* 137 (2023), 109347.

[50] Wang Yin, Zhao Jingyong, Xie Gang, Zhao Zhicheng, and Hu Xiao. 2024. PA-YOLO-Based Multifault Defect Detection Algorithm for PV Panels. *International Journal of Photoenergy* 2024, 1 (2024), 6113260.

[51] Hengfu Yu, Jinhong Deng, Wen Li, and Lixin Duan. 2024. Towards Unsupervised Model Selection for Domain Adaptive Object Detection. *Advances in Neural Information Processing Systems* 37 (2024), 58423–58444.

[52] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. 2024. Bilateral Reference for High-Resolution Dichotomous Image Segmentation. *CAAI Artificial Intelligence Research* (2024).

[53] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2022. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence* 45, 4 (2022), 4396–4415.

[54] Orr Zohar, Alejandro Lozano, Shelly Goel, Serena Yeung, and Kuan-Chieh Wang. 2023. Open world object detection in the era of foundation models. *arXiv preprint arXiv:2312.05745* (2023).