Lightweight Thermal Super-Resolution and Object Detection for Robust Perception in Adverse Weather Conditions

Pranjay Shyam
Faurecia IRYStec Inc.
Montreal, Canada
pranjay.shyam.psm@forvia.com

HyunJin Yoo
Faurecia IRYStec Inc.
Montreal, Canada
hyunjin.yoo@forvia.com

Abstract

In this work, we examine the potential application of thermal cameras in improving perception capabilities in adverse weather conditions like snow, night-time driving, and haze, focusing on retaining the performance of Advanced Driver Assistance Systems (ADAS), thus enhancing its functionality and safety characteristics. While thermal sensors offer the advantage of robust information capture in adverse weather conditions, their integration is plagued with issues surrounding poor feature capture in normal conditions, low imaging resolution, and high sensor costs. We address the former by formulating the problem definition as information switching wherein thermal images are selected when visible images are degraded. Furthermore, we consider a single object detector for RGB and thermal images to ensure low latency. We propose utilizing a learnable projection function that translates the thermal image into RGB color space, thus providing minimal modifications to the underlying object detector. We address the issues of low imaging resolution and cost by proposing a novel procedure that combines super-resolution and object detection, enabling the utilization of low-resolution and low-cost uncooled thermal imaging sensors. To ensure the complete pipeline meets the actual deployment requirements of real-time inference on resource-constrained devices, we introduce a lightweight super-resolution algorithm, implementing optimizations within the network structure followed by global pruning. In addition, to improve the feature representations extracted by lightweight encoders, we propose a bidirectional feature pyramid network to enhance the feature representation. We demonstrate the efficacy of the proposed mechanism through extensive simulated evaluations on automotive datasets such as FLIR, KAIST, DENSE, and Freiburg Thermal.

1. Introduction

Adverse weather conditions, such as fog, rain, snow, and low illumination, pose significant challenges to the perception stack of ADAS systems. Since these systems primarily depend on RGB images, poor information captured in such conditions results in the failure of underlying perception tasks such as semantic segmentation [25, 56, 62, 66, 73, 85], object detection [28, 37, 45, 55, 64, 67, 68, 86], depth estimation [80] and simultaneous localization and mapping [26]. While general purpose image restoration algorithms [38, 54, 62, 74] can be leveraged to improve the image quality, their domain specific nature [63] restricts their practicality.

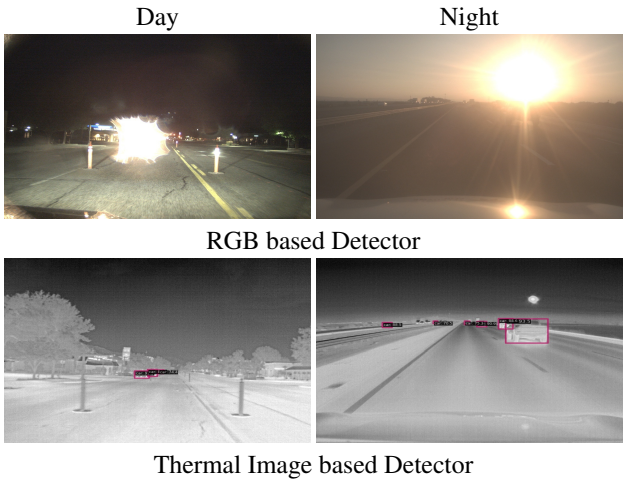


Figure 1. Results on day and night images from FLIR [2] dataset.

To address this issue and improve driver situational awareness under adverse weather conditions, we propose the integration of thermal infrared (IR) sensors, as these remain mainly unaffected by external illumination and environmental factors, making them an ideal solution. In contrast to traditional IR imaging systems that necessitate external illumination sources [5], thermal imaging systems capture the inherent thermal radiation emitted by the surface of objects. This characteristic is especially beneficial when visual degradation hinders human and machine perception, such as dealing with headlight flare during nighttime driving [16] or driving in foggy conditions [5]. Hence, these

sensors can serve a crucial role in improving the reliability of ADAS systems in adverse weather conditions.

Despite the inherent advantages of thermal imaging sensors, their integration into the perception stack of vehicles has been hindered by issues such as poor information capture in clear conditions, high sensor cost, and low image resolution. Nevertheless, their value addition in terms of safety makes their incorporation crucial. In contrast to prior fusion-based approaches [59, 89, 91], we present a novel switching mechanism that seamlessly integrates thermal cameras into the perception stack with minimal modifications. The proposed mechanism utilizes thermal images as input to the underlying object detection algorithm when RGB images are severely corrupted, ensuring robust perception in challenging environmental conditions. To ensure performance retention of the object detector towards change in modalities (thermal and RGB), we propose a learnable projection function that effectively transforms information from the thermal domain into the RGB domain, preserving the performance of the underlying object detector while facilitating efficient processing and decision-making within the perception stack.

While high-performance thermal imaging sensors open for consumer applications tend to be expensive, these are still restricted to a maximum resolution of 640 $\times$512 pixels, which is low compared to RGB sensors, where the maximum pixel resolution reaches up to 8MP. However, there exist cheaper micro-thermal cameras [1] that can be used as an alternative, albeit with a lower resolution of 160 $\times$ 120 pixels. Motivated by such sensors and their future application in the automotive domain, we propose a software-based framework for their integration to demonstrate robustness ensured in adverse weather conditions. However, their reduced resolution poses a challenge to such integration. Thus, to overcome this, we propose a lightweight super-resolution algorithm that provides a high-resolution thermal image from its low-resolution counterpart. Enabling the same thermal and RGB image resolution allows us to utilize either for an underlying object detector. Since such a framework is to be deployed in automotive applications, computer restrictions apply. While prior works utilize a compute-efficient backbone for extracting features that are used for object detection, the lightweight nature of such algorithms results in reduced feature quality. Thus, to improve the feature quality subsequently used for localizing and classifying objects of interest, we propose modification into the feature pyramid network wherein we propose a compute efficient bi-directional FPN. Furthermore, we utilize the designed object detection during the optimization of the underlying super-resolution network to ensure consistent super-resolution without adversarial patterns. To evaluate the efficacy of the proposed mechanism, we construct a segment-anything [35] extension for thermal images to generate pseudo training labels for several datasets such as FLIR [2], KAIST [10], DENSE [5], and Freiburg Thermal [75]. We elaborate upon the problem setting corresponding to these datasets in Sec. 3 and summarize our methodologies as,

- We propose a switching mechanism that ensures dynamic utilization of thermal images in the event of poor quality RGB images.
- To ensure compatibility between the thermal and RGB modalities, we propose a projection function that transforms the thermal image into compatible RGB space.
- To utilize low-cost, low-resolution thermal images, we propose a lightweight super-resolution algorithm that follows hardware-aware efficient design.
- To improve the feature information captured by lightweight feature extraction backbones, we propose a bi-directional feature pyramid network.
- We utilize the object detection algorithm during optimization to reduce adversarial patterns in the super-resolved images.
- We propose an extension to the Segment-Anything model for thermal images to generate 2D bounding boxes for training and evaluation purposes.

2. Related Works

2.1. Thermal Sensors in Automotive Perception

In recent years, numerous approaches have been proposed to leverage thermal information to enhance the performance of perception algorithms in adverse weather and illumination conditions. Some early works focused on image fusion [88] or feature enhancement [72] to improve object detection [3, 37, 40, 53] algorithms. Among the early works, the KAIST dataset [10] was introduced to explore the influence of different sensor modalities, aiming to achieve all-weather perception capabilities. Prior researchers [88] focused on examining the possibility of fusing RGB and Thermal images to ensure consistent performance of object detection algorithms. Alternatively, [12] evaluated utilizing GANs to convert thermal images into RGB space for improved object detection performance was explored. However, it was noted that directly fusing features across modalities generates sub-optimal results. With this motivation, different attention mechanisms [59, 89, 91] were proposed to improve feature quality for optimal fusion and detection. Alternatively, researchers also focused on improving the attributes that can be extracted from thermal images restricted to 2D object detection. Towards this objective [19, 33, 75] focused on performing semantic segmentation on thermal images without utilizing prior annotated data by following the approach of unsupervised domain adaption.

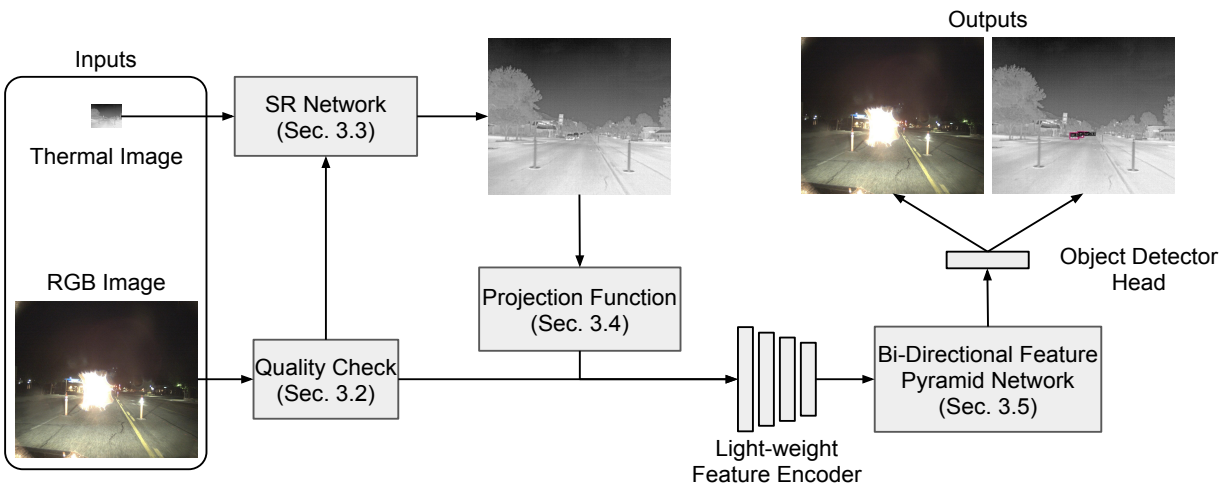


Figure 2. Overview of the proposed framework to ensure consistent object detection performance in events of adverse weather conditions.

2.2. Real-Time Object Detection

From the aspect of automotive deployment, real-time object detection is desired with object detection landscape being divided into anchor [6, 23, 58, 65, 76, 77] based and anchor free [20, 24, 39, 87] categories. While anchor-based object detectors were widely used, their reliance on expensive post-processing operation of non-maximum suppression (NMS) resulted in performance bottlenecks. Current approaches for real-time object detection focus on lightweight encoder [27, 60, 78], improving feature representation via feature pyramid network [22, 65], lightweight object detection head [20, 87] and elimination of complex post-processing non-maximum suppression. Recently end-to-end object detectors based on transformer architecture were proposed (DETR [7]), wherein the need for expensive NMS was eliminated by using bipartite matching. Subsequent versions of DETR focused on improving training convergence [9, 51, 84, 94]. As current SoTA real-time object detectors rely on lightweight feature extractors, we emphasize on improving the feature quality by proposing a bi-directional feature enhancement network. While prior works [32, 43, 71, 79, 93] proposed similar mechanisms, we highlight and address the performance bottleneck arising from inefficiencies of convolution-based multi-level feature fusion.

2.3. Lightweight Super Resolution

The growing applications of super-resolution algorithms resulted in increased interest in improving efficiency to deploy such algorithms in real-time on resource-constrained devices [4, 11, 15, 18, 29, 36, 41, 44, 81]. Towards this objective, solutions such as reducing convolutional kernel size [14], utilizing cascaded residual blocks [4]. To further improve the efficiency of the underlying network architecture, [29] proposed a multi-distillation block, which was improved by [44] that utilized a residual feature distillation block. Recently [13] proposed expansion of optimization space using a structural reparameterization technique wherein a multi-branch training network is simplified to a feed-forward inference network. Given these efficient super-resolution networks, remote sensing deployments were identified to improve object detection performance on satellite imagery [17,31,52,57,61,83,90,95]. Unlike prior works, we extend the scope of object-detection-guided super-resolution to thermal images, with the deployment scenario focusing on improving driving decisions in adverse weather conditions. Such a scenario requires real-time inference on resource-constrained devices and is not explored by prior works.

3. Methodology

3.1. Problem Overview

Given a low-resolution thermal image (I^T_{LR}) and an aligned high-resolution RGB image (I_{RGB}), the objective is to leverage the thermal image in the event of poor RGB quality to be used as input to the object detection algorithm. Towards this, a light-weight super-resolution network is proposed to upsample (x4) the thermal image (I^T_{HR}), which is subsequently projected into RGB space for utilization by the shared object detector.

3.2. Switching Functionality

The demand for lightweight and computationally efficient solutions is paramount in the realm of image quality assessment for optimal performance of object detection algorithms. Addressing this need, we propose a pragmatic approach to swiftly detect potential image degradation before integrating the input into the underlying object detection pipeline. Focusing on weather-induced corruptions, including flare, snow, fog, rain, and low-light scenarios, we observe that these often lead to over-saturation, creating regions of intense brightness within the image. Capitalizing on this insight, we introduce a straightforward yet effective strategy. By designing a basic filter, we identify dis-

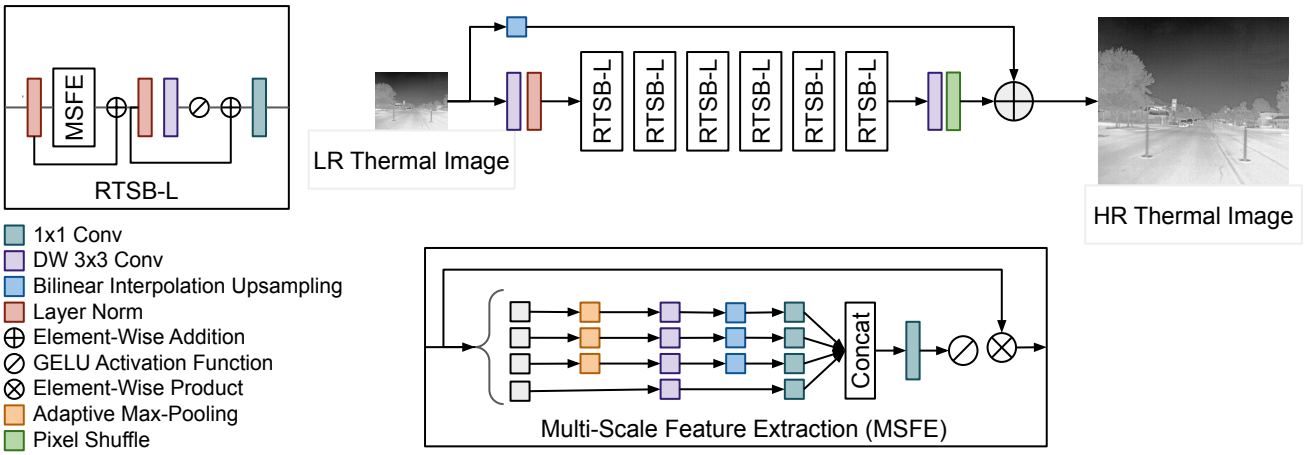


Figure 3. Overview of the proposed lightweight super resolution framework.

tinctive peaks in the intensity histogram of the RGB image. These peaks, indicative of the predominant intensity values, are averaged to determine the highest peak. Employing a thresholding mechanism (β_1, β_2), we can classify whether the given image has been compromised by weather-induced degradation. Our proposed solution offers a simple means of preemptively assessing image quality, thus ensuring the robustness and accuracy of downstream object detection tasks, even under challenging weather conditions. We present the corresponding algorithm in Algo. 1 with qualitative samples in Appendix-A of supplementary. From empirical evaluation we fix β_1 and β_2 to 180 and 20 respectively.

Algorithm 1 Weather Degradation Identification

Input $\leftarrow$ RGB image
Convert RGB image to grayscale using luminance conversion.
Apply median filter to reduce noise
Compute image intensity histogram
Find the peaks in the histogram
Sort the peaks in ascending order
Compute the average intensity value of the highest peak
Set a threshold for weather (β_1) and Illumination degradation β_2
if average intensity value $\geq \beta_1$ **then**
    Return $\rightarrow$ Weather Degradation
else if average intensity value $\leq \beta_2$ **then**
    Return $\rightarrow$ Illumination Degradation
else
    Return $\rightarrow$ No Significant Weather Degradation
end if

3.3. Light-Weight Super Resolution

Given the presence of weather degradation affecting RGB image quality, we perform lightweight super-resolution for the thermal image. We adopt a two-fold strategy to build upon the insights drawn from [48], which underscored the direct correlation between deep learning network inference speed and activation volume. Initially, we streamline the computational load of the super-resolution network by tactically downsampling the input image across the channel dimension through strided convolution operations. This strategic spatial compression enhances computational efficiency and serves as a precursor for our subsequent innovations. Recognizing the inherent ill-posed nature of low-resolution super-resolution problems [47–49], wherein multiple high-resolution solutions coexist, we harness the prowess of transformer-based super-resolution architectures for adeptly capturing intricate non-local feature correlations.

Nonetheless, the computational demands of transformer models are notorious, largely due to the quadratic complexity stemming from self-attention mechanisms. Addressing this hurdle, we introduce a novel convolution-based alternative to conventional self-attention. This innovative approach facilitates the extraction of multi-scale features that subsequently undergo dynamic feature selection. This dynamic selection technique ensures the assimilation of non-local feature interactions, which are then synergistically augmented with the power of convolutional channel mixture strategies [70]. This synergy enables the efficient extraction of pertinent local features. Collectively, our lightweight yet holistic framework emerges as a compelling solution for robust thermal image super-resolution under the complex influence of diverse weather conditions while simultaneously adhering to the imperative of computational efficiency for real-world applications.

In addressing the crucial need for capturing long-range dependencies while circumventing the computational challenges posed by self-attention mechanisms, we present an innovative solution rooted in feature pyramid networks (FPNs) tailored to cater to diverse scales of contextual information. Our approach entails a multi-step process that

effectively marries global and local feature integration. To elaborate, we kickstart the process by constructing a robust FPN architecture leveraging channel splitting techniques, thereby facilitating the extraction of multi-scale features across four distinct scales (1, 1/2, 1.4, and 1/8). These diverse scale-specific features are then refined through a judicious combination of operations that balance information preservation and computational efficiency.

In particular, a 3×3 depth-wise convolution is a pivotal element, allowing us to channel the extracted features into a transformative phase. This is followed by an adaptive nearest interpolation to homogenize feature dimensions across the scales. To amplify the richness of the fused features, we employ a refined 1×1 convolution, imparting the necessary enhancement while preserving the computational economy paramount for real-world applicability. Significantly, our approach incorporates a GELU activation function, acting as an enabler for introducing non-linearity, thereby fostering the intricate representations that are quintessential for robust feature extraction. This holistic methodology yields a feature-rich representation that encapsulates global context and fine-grained local information, all while circumventing the traditionally associated quadratic complexities of the self-attention mechanism.

We replace the standard feed-forward layer with a convolutional channel mixer [70] to enhance further the local spatial modeling within the modified transformer block. Unlike prior works that proposed utilizing 1×1 convolutions or fully connected layer, the alternative mechanism uses 3×3 convolutions to expand features across channel dimensions followed by mixing operation. Finally, 1×1 convolution is applied to compress the feature space. We present an overview of proposed super-resolution algorithm in Fig. 3.

3.4. Learnable Projection Function

In the realm of super-resolution for thermal images, we confront the formidable challenge of executing robust object detection tasks while treading lightly on the computational front. Prior endeavors in this domain have typically resorted to employing distinct object detectors tailored to different modalities or adopting a concatenation strategy that combines RGB and thermal imagery. However, we steer our approach in a different direction. We propose a solution to integrate thermal imagery into the object detection pipeline seamlessly. We have introduced a learnable projection function that orchestrates the transformation of thermal image statistics into a format that seamlessly aligns with the RGB space. This operation is grounded in channel-wise mean-variance transfer. Mathematically, this novel projection is succinctly expressed as:

$$I_{\widehat{RGB}} = \frac{I^T_{HR} - \mu_{I^T_{HR}}}{\sigma_{I^T_{HR}}} \cdot \sigma + \mu \qquad (1)$$

here I^T_{HR} represents the input high resolution thermal image, while $\mu_{I^T_{HR}}$ and $\sigma_{I^T_{HR}}$ correspond to its mean and standard deviation, respectively. For each channel in the R, G, B space σ, μ are learnable parameters. To ensure that optimizing these parameters does not break the training cycle due to boundary conditions when $\sigma = 0$ we include a fixed bias of 1e-3.

The resultant $I_{\widehat{RGB}}$ is an image that faithfully encapsulates the thermal characteristics within the RGB domain. By performing this translation in statistics, we pave the way for employing a single, unified object detector—a detector that remains invariant to the input modality. This groundbreaking approach not only simplifies the computational complexity but also elevates the robustness and versatility of our object detection system.

3.5. Bi-Directional Feature Pyramid Network

We propose modifications within the feature pyramid network to improve the feature representation within the single-stage object detectors while requiring less computation. Specifically, we propose performing 1×1 convolution on multi-scale features before upsampling, resulting in the same output resolution as traditional FPN while consuming fewer parameters and floating point operations. Second, we refine the multi-scale features using group-wise 1×1 convolution. Finally, instead of using multi-scale features for performing object detection, following prior real-time object detection algorithms [20, 23, 24, 50], we use the aggregated feature map to compute bounding box for objects of interest. We refer to this mechanism as a bi-directional feature pyramid network and present an overview in Fig. 4.

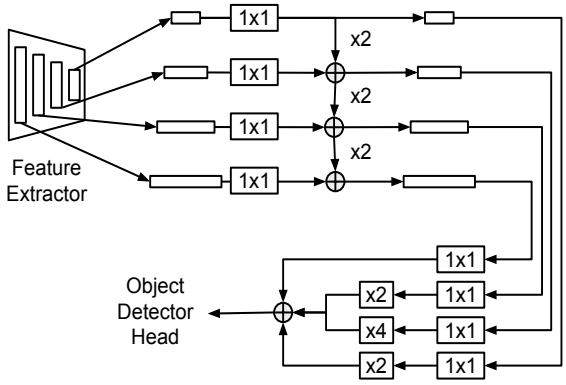


Figure 4. Overview of the proposed bi-directional feature pyramid network.

3.6. Training Mechanism

We follow a three-stage training pipeline wherein we train the object detector first on RGB images from the FLIR dataset while integrating the proposed bi-directional FPN. Since FPN is a common component within SoTA real-time object detectors, it can be integrated easily within any framework. For our experiments, we consider RTMDet [50]

built on over mmdetection [8]. We train the aforementioned object detection using AdamW [34] ($\beta_1 = 0.9, \beta_2 = 0.999$) optimizer with a learning rate of $4e^{-3}$ and a weight decay of 0.05 following a cosine annealing [46] learning rate strategy for 300 epochs at an input resolution of 640 $\times$ 512.

Secondly, we train the underlying super-resolution algorithm alongside the projection function parameters while keeping the object detector fixed. For these settings, we train for 1000 epochs with a learning rate of $1e^{-3}$ adjusted via cosine annealing [46] to $1e^{-5}$ and ADAM [34] optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$). For loss computation, we follow prior works and utilize a combination of L_1 and weighted FFT loss [64] following,

$$L = ||SR(I^T_{LR}) - I^T_{HR}||_1 + \lambda * ||FFT(SR(I^T_{LR})) - FFT(I^T_{HR})||_1$$

Here λ represents the weight parameter and is fixed to 0.1 based on empirical evaluation. Finally, we train the complete framework for 500 epochs using a mix of RGB and Thermal images from the FLIR dataset with a learning rate of $1e^{-5}$ adjusted via cosine annealing to $1e^{-7}$ using ADAM optimizer and combine the loss function for object detection and super-resolution.

4. Experimental Evaluation

4.1. Datasets and Evaluation Metrics

For our experiments, we use publicly available datasets such as FLIR [2], KAIST [10], DENSE [5], and Freiburg Thermal [75]. Among these, the FLIR dataset stands out as a particularly comprehensive benchmark, boasting a diverse array of 15 distinct classes and a voluminous dataset encompassing 9711 thermal images and 9233 RGB images, each boasting a resolution of 640 $\times$ 512. In contrast, the KAIST dataset offers annotations specifically tailored to pedestrians, a feature that sets it apart. To fortify our experimental setup, we augment the KAIST, DENSE, and Freiburg Thermal datasets to incorporate an expanded array of label attributes. This augmentation endeavors to encompass a broader spectrum of objects, spanning categories such as Persons, Bikes, Cars, Motorcycles, Buses, Trains, Trucks, Traffic Lights, Fire Hydrants, Street Signs, Dogs, Skateboards, Strollers, Scooters, and Other Vehicles. We execute this augmentation by strategically combining the Segment Anything Model and the Domain-Adaptive Panoptic Segmentation technique. Our augmentation scheme's intricacies, alongside detailed insights into class distribution and resolution variations, are meticulously elaborated upon in Appendix B of the supplementary materials. To simulate the lower-resolution thermal imagery, we adhere to established conventions by applying the bicubic downsampling algorithm, supplemented by noise utilizing noise models presented in the SCUNet framework. We draw upon the widely adopted PSNR and SSIM metrics for quantitative performance assessments to gauge the quality of super-resolution results. Additionally, to assess object detection performance, we rely on the mAP (@0.50 IoU) metric, renowned for its comprehensive evaluation of detection efficacy.

4.2. Comparison with SoTA : Thermal Super Resolution

We perform comprehensive evaluation encompassing both qualitative performance assessment along side investigating the computational complexity measured using floating-point operations per second (FLOPs) and the overall parameter count. Towards this evaluation, we benchmark the performance of state-of-the-art (SOTA) lightweight super-resolution methodologies for thermal super resolution and summarize the quantitative results in Table 2. To ensure fair evaluation, we retrain these algorithms adhering to the hyperparameters stipulated by the authors of corresponding papers. This retraining is conducted leveraging the low-resolution images generated by the method delineated in Section 4.1. Based on the qualitative and quantitative assessments, we conclude the proposed super resolution algorithm to surpass SoTA performance while relying on less computational resources in terms of lower parameter count and reduced computational complexity. Notably, our optimization strategy takes two distinct paths: one relies solely on high-resolution thermal images, while the other harnesses the synergistic insights provided by the integrated object detector. In subsequent evaluation, we distinguish these two optimization paths as "Ours-I" and "Ours-II," respectively. An in-depth perusal of Table 2 underscores the superiority of our proposed super-resolution algorithm (Ours-I), surpassing the established ShuffleMixer [69] while simultaneously maintaining a commendably lean computational footprint and parameter requirement. Intriguingly, we also observe a modest yet consistent performance improvement when our super-resolution algorithm is jointly optimized with the object detector. This improvement can be attributed to the introducing of a systematic bias from the object detector, thereby endowing the super-resolution algorithm with improved structural precision. It is worth noting that this introduction of bias is not indicative of dataset leakage, as both the object detector and the super-resolution algorithm are trained on the same training dataset, ensuring a uniform and equitable training environment. For completeness, we summarize performance when using SoTA super resolution trained alongside object detection mechanism following proposed mechanism in Tab. 1.

4.3. Comparison with State-of-the-Art: Real-Time Object Detection

To evaluate the performance of our real-time object detection baseline and its variants enhanced with our proposed

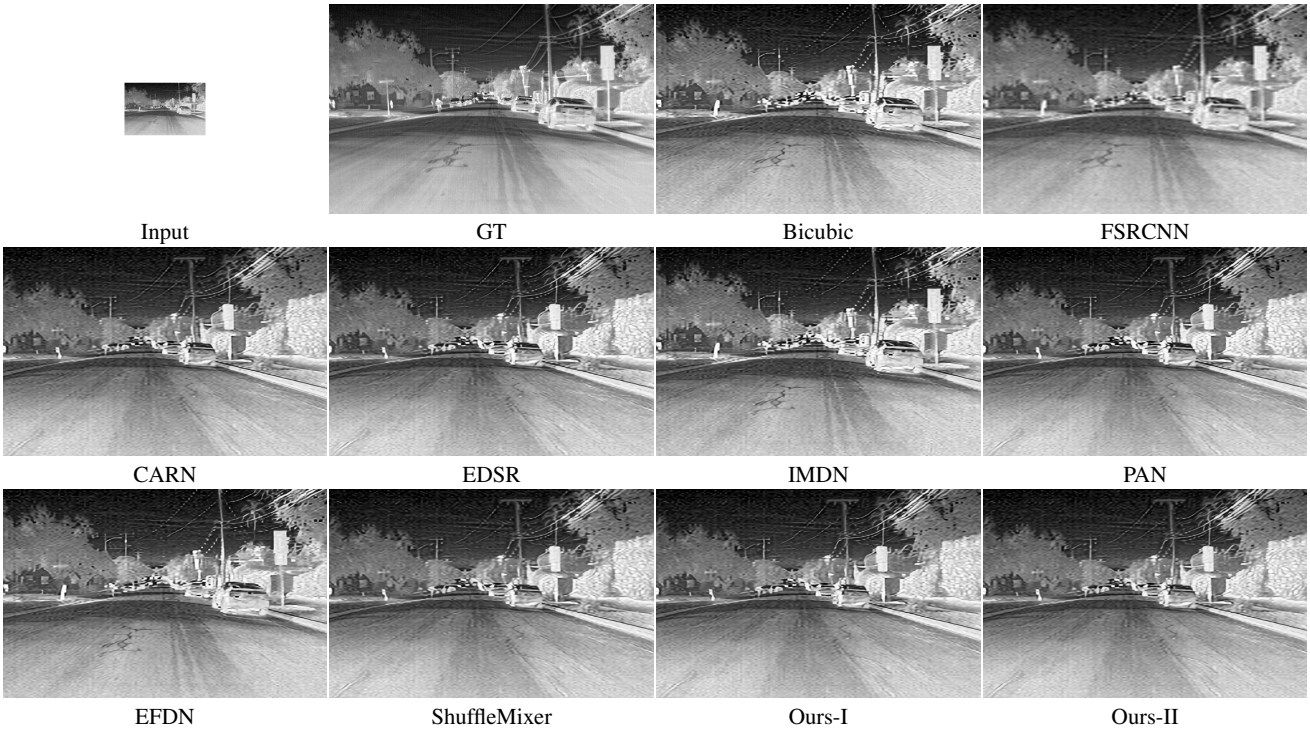


Figure 5. Qualitative performance evaluation of SoTA super-resolution algorithms retrained on FLIR dataset [2]

bi-directional feature pyramid network, we conducted comprehensive comparisons. We extended these object detection algorithms by seamlessly integrating them with our super-resolution algorithm, as detailed in Appendix-D of our supplementary materials. Our primary focus was on compact versions of established frameworks, specifically RTMDet [50] and YOLOX [21]. Our evaluation involved a meticulous analysis, and the quantitative results are provided in Table 3. Throughout this evaluation, our training process adhered closely to the prescribed methodologies of the original frameworks, maintaining the specified hyper-parameters. Notably, when training with thermal images, we adjusted the number of input channels to a single entity. We observed a significant performance improvement for both RTMDet and YOLOX, affirming the effectiveness of our proposed enhancements. Importantly, this performance boost was achieved without compromising the efficient allocation of parameters and computational operations. Our empirical investigation underscores the enduring potential of our enhancements, making them versatile tools capable of delivering robust object detection results in both RGB and thermal imaging domains.

Table 1. Influence of integrating object detection algorithm (YOLOX-tiny detection) during optimization on performance of SoTA super resolution algorithms using DENSE and FLIR datasets.

SR-Method	DENSE	FLIR	FLIR-Thermal Detection
Bicubic	25.15 / 0.21	25.92 / 0.70	47.28
FSRCNN [14]	25.73 / 0.65	26.97 / 0.73	48.18
CARN [4]	25.56 / 0.71	26.45 / 0.77	49.51
EDSR [42]	25.40 / 0.75	25.98 / 0.78	49.57
IMDN [30]	25.07 / 0.66	25.75 / 0.77	49.82
PAN [92]	25.46 / 0.76	26.01 / 0.79	49.24
EFDN [82]	26.02 / 0.72	26.28 / 0.73	49.22
ShuffleMixer [69]	26.04 / 0.77	26.36 / 0.78	48.98

Table 2. Quantitative performance of SoTA super resolution algorithm for DENSE and FLIR datasets.

Method	DENSE	FLIR	# Params(K)	FLOPs(G)
	PSNR / SSIM	PSNR / SSIM		
Bicubic	25.15 / 0.21	25.92 / 0.70	-	-
FSRCNN [14]	25.54 / 0.64	26.46 / 0.70	12	5
CARN [4]	25.47 / 0.61	26.44 / 0.70	1503	87
EDSR [42]	25.31 / 0.66	25.97 / 0.70	1498	107
IMDN [30]	24.96 / 0.65	25.69 / 0.69	695	39
PAN [92]	25.39 / 0.66	26.01 / 0.70	257	21
EFDN [82]	25.95 / 0.21	26.26 / 0.70	1051	41
ShuffleMixer [69]	26.18 / 0.65	26.52 / 0.72	398	28
Ours-I	26.24 / 0.62	26.48 / 0.72	220	12
Ours-II	26.95 / 0.67	27.16 / 0.75	220	12

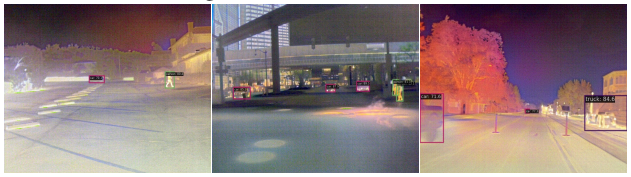
4.4. Ablation Studies

4.4.1 Influence of modality change

A systematic exploration into the profound implications of modality alterations on the foundational object detection paradigm underscores the crux of our analysis. This extensive inquiry is undertaken by scrutinizing the performance of an RGB image-oriented object detector when confronted with thermal images and vice versa. Our investigation is

Thermal Image on RGB trained detector

RGB Image on Thermal trained detector



Proposed Methodology

Figure 6. Qualitative examples demonstrating change in input modality on pretrained object detection algorithm.

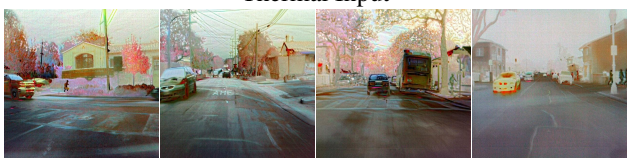
meticulously carried out to evaluate the impact on detection capabilities, which is pivotal for understanding the interplay between different sensory modalities. As part of this analysis, we venture into the domain of thermal imaging, facilitating the translation of single-channel thermal images to RGB space through channel-wise duplication. While necessitated by input dimension compatibility, this approach yields a marginal mAP of 26.70, starkly contrasting with the baseline performance benchmark of 70.91.

Table 3. Quantitative performance of SoTA real time object detection algorithms on FLIR dataset.

Method	FLIR-RGB	FLIR-Thermal	FLOPs (G)	# Params (M)
YOLOX-tiny [21]	58.51	57.08	6.3	4.9
RTMDET-tiny [50]	67.83	67.19	7.8	4.3
YOLOX-tiny-Ours	61.29	60.59	6.1	4.6
RTMDET-tiny-Ours	70.91	69.84	7.2	4.0

Conversely, we delve into the consequences of deploying RGB images as input for a thermal image-trained object detector. This pursuit entails converting RGB images to grayscale before inferencing them using the object detection algorithm. This approach manifests a corresponding map of 32.14, diverging remarkably from the baseline performance of 69.84. Fig. 6 encapsulates the crux of these explorations in a succinct visual representation. These profound observations unveil that modality transition instigates a precipitous decline in detection performance, echoing the underlying incompatibility intrinsic to the image space. In response to this foundational challenge, we present a novel solution that reflects our proposed methodology. As a testament to the robustness and efficacy of this mechanism, we achieved an outstanding baseline mAP of 29.41 across the comprehensive FLIR dataset. This performance boost is accompanied by a tangible enhancement in detection confidence and the compactness of bounding boxes in the projected thermal space, a revelation underscored through a meticulous evaluation of qualitative outcomes. For the sake of comprehensiveness, we augment our discourse by providing additional insights within Appendix C of the supplementary materials. We provide qualitative example of conversion of thermal image to RGB image using proposed projection function in Fig. 7.

Thermal Input



Projected RGB Image

Figure 7. Conversion of thermal image (top) to RGB image (bottom) compatible with object detector using the proposed projection function.

5. Conclusion

With the objective to retain performance of image based perception systems in adverse weather and illumination conditions, we design a framework integrating information from a low-cost low-resolution thermal camera alongside high resolution RGB image. In order to integrate thermal image in adverse weather conditions, we first perform a quality check for identification of saturated regions or low illumination conditions. In event of poor image quality, we utilize thermal image as the input for the object detection algorithm. Since the resolution and modality of thermal image is different from RGB image, we incorporate a lightweight super resolution network to upsample the thermal image by a scale of 4. To address the modality change, we propose a learnable projection function which maps the super resolved thermal image into RGB space. Finally the image is processed by a light weight object detection algorithm to identify objects of interest. We perform extensive studies to evaluate the performance of proposed mechanism vis-a-vis combinations involving use of RGB or Thermal images only and obtain superior results compared to using either. Importantly, our approach offers computational efficiency, making it suitable for real-time applications.

References

[1] Lwir micro thermal camera module. https://www.flir.com/products/lepton/?vertical=microcam&segment=oem/ [Accessed: (Feb. 22, 2023)]. 2

[2] Flir thermal dataset, 2021. https://www.flir.com/oem/adas/adas-dataset-form/ [Accessed: (Dec. 22, 2022)]. 1, 2, 6, 7

[3] Kshitij Agrawal and Anbumani Subramanian. Enhancing object detection in adverse conditions using thermal imaging. *arXiv preprint arXiv:1909.13551*, 2019. 2

[4] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. *arXiv preprint arXiv:1803.08664*, 2018. 3, 7

[5] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11682–11692, 2020. 1, 2, 6

[6] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 3

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 3

[8] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[9] Qiang Chen, Xiaokang Chen, Gang Zeng, and Jingdong Wang. Group detr: Fast training convergence with decoupled one-to-many label assignment. *arXiv preprint arXiv:2207.13085*, 2022. 3

[10] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948, 2018. 2, 6

[11] Marcos V Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte. Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration. In *Proceedings of Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 669–687. Springer, 2023. 3

[12] Chaitanya Devaguptapu, Ninad Akolekar, Manuj M Sharma, and Vineeth N Balasubramanian. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[13] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13733–13742, 2021. 3

[14] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 391–407. Springer, 2016. 3, 7

[15] Zongcai Du, Ding Liu, Jie Liu, Jie Tang, Gangshan Wu, and Lean Fu. Fast and memory-efficient network towards efficient image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 853–862, 2022. 3

[16] Anil Erkan, David Hoffmann, Timo Singer, Julia Maria Schikowski, Korbinian Kunst, Markus Alexander Peier, and Tran Quoc Khanh. Influence of headlight level on object detection in urban traffic at night. *Applied Sciences*, 13(4):2668, 2023. 1

[17] Syeda Nyma Ferdous, Moktari Mostofa, and Nasser M Nasrabadi. Super resolution-assisted deep aerial vehicle detection. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, pages 432–443. SPIE, 2019. 3

[18] Guangwei Gao, Wenjie Li, Juncheng Li, Fei Wu, Huimin Lu, and Yi Yu. Feature distillation interaction weighting network for lightweight image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 661–669, 2022. 3

[19] Huan Gao, Jichang Guo, Guoli Wang, and Qian Zhang. Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9913–9923, 2022. 2

[20] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 3, 5

[21] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 7, 8

[22] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7036–7045, 2019. 3

[23] Jocher Glenn. Yolov5 release v7.0. https://github.com/ultralytics/yolov5/tree/v7.0, 2022. 3, 5

[24] Jocher Glenn. Yolov8. https://github.com/ultralytics/ultralytics/tree/main, 2023. 3, 5

[25] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference*

on Intelligent Robots and Systems (IROS), pages 5108–5115. IEEE, 2017. 1

[26] Sungchul Hong, Pranjay Shyam, Antyanta Bangunharcana, and Hyuseoung Shin. Robotic mapping approach under illumination-variant environments at planetary construction sites. *Remote Sensing*, 14(4):1027, 2022. 1

[27] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 3

[28] Kaiqi Huang, Liangsheng Wang, Tieniu Tan, and Steve Maybank. A real-time object detecting and tracking system for outdoor night surveillance. *Pattern Recognition*, 41(1):432–444, 2008. 1

[29] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2024–2032, 2019. 3

[30] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th ACM International Conference on Multimedia (ACM MM)*, pages 2024–2032, 2019. 7

[31] Hong Ji, Zhi Gao, Tiancan Mei, and Bharath Ramesh. Vehicle detection in remote sensing images leveraging on simultaneous super-resolution. *IEEE Geoscience and Remote Sensing Letters*, 17(4):676–680, 2019. 3

[32] Seung-Wook Kim, Hyong-Keun Kook, Jee-Young Sun, Mun-Cheon Kang, and Sung-Jea Ko. Parallel feature pyramid network for object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 234–250, 2018. 3

[33] Yeong-Hyeon Kim, Ukcheol Shin, Jinsun Park, and In So Kweon. Ms-uda: Multi-spectral unsupervised domain adaptation for thermal image semantic segmentation. *IEEE Robotics and Automation Letters*, 6(4):6497–6504, 2021. 2

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[35] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2

[36] Fangyuan Kong, Mingxi Li, Songwei Liu, Ding Liu, Jingwen He, Yang Bai, Fangmin Chen, and Lean Fu. Residual local feature network for efficient super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 766–776, 2022. 3

[37] Mate Krišto, Marina Ivasic-Kos, and Miran Pobar. Thermal object detection in difficult weather conditions using yolo. *IEEE access*, 8:125459–125476, 2020. 1, 2

[38] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17452–17462, 2022. 1

[39] Chuyi Li, Lulu Li, Yifei Geng, Hongliang Jiang, Meng Cheng, Bo Zhang, Zaidan Ke, Xiaoming Xu, and Xiangxiang Chu. Yolov6 v3. 0: A full-scale reloading. *arXiv preprint arXiv:2301.05586*, 2023. 3

[40] Shasha Li, Yongjun Li, Yao Li, Mengjun Li, and Xiaorong Xu. Yolo-firi: Improved yolov5 for infrared image object detection. *IEEE access*, 9:141861–141875, 2021. 2

[41] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 3

[42] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 7

[43] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[44] Jie Liu, Jie Tang, and Gangshan Wu. Residual feature distillation network for lightweight image super-resolution. In *Proceedings of Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 41–55. Springer, 2020. 3

[45] Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive yolo for object detection in adverse weather conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1792–1800, 2022. 1

[46] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[47] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Ntire 2021 learning the super-resolution space challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 596–612, 2021. 4

[48] Andreas Lugmayr, Martin Danelljan, Radu Timofte, Kangwook Kim, Younggeun Kim, Jae-young Lee, Zechao Li, Jinshan Pan, Dongseok Shim, Ki-Ung Song, et al. Ntire 2022 challenge on learning the super-resolution space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 786–797, 2022. 4

[49] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 715–732. Springer, 2020. 4

[50] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors, 2022. 5, 7, 8

[51] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3651–3660, 2021. 3

[52] Moktari Mostofa, Syeda Nyma Ferdous, Benjamin S Riggan, and Nasser M Nasrabadi. Joint-srvdnet: Joint super resolution and vehicle detection network. *IEEE Access*, 8:82306–82319, 2020. 3

[53] Farzeen Munir, Shoaib Azam, Muhammd Aasim Rafique, Ahmad Muqeem Sheri, Moongu Jeon, and Witold Pedrycz. Exploring thermal images for object detection in underexposure regions for autonomous driving. *Applied Soft Computing*, 121:108793, 2022. 2

[54] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–12, 2023. 1

[55] Heena Patel and Kishor P Upla. Night vision surveillance: Object detection using thermal and visible images. In *2020 International Conference for Emerging Technology (INCET)*, pages 1–6. IEEE, 2020. 1

[56] Andreas Pfeuffer and Klaus Dietmayer. Robust semantic segmentation in adverse weather conditions by means of sensor data fusion. In *2019 22th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2019. 1

[57] Jakaria Rabbi, Nilanjan Ray, Matthias Schubert, Subir Chowdhury, and Dennis Chao. Small-object detection in remote sensing images with end-to-end edge-enhanced gan and object detector network. *Remote Sensing*, 12(9):1432, 2020. 3

[58] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 3

[59] A Sai Charan, M Jitesh, M Chowdhury, and H Venkataraman. Abifn: Attention-based bi-modal fusion network for object detection at night time. *Electronics Letters*, 56(24):1309–1311, 2020. 2

[60] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 3

[61] Jacob Shermeyer and Adam Van Etten. The effects of super-resolution on object detection performance in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[62] Pranjay Shyam, Kyung-Soo Kim, and Kuk-Jin Yoon. Giqe: Generic image quality enhancement via nth order iterative degradation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2077–2087, 2022. 1

[63] Pranjay Shyam, Sandeep Singh Sengar, Kuk-Jin Yoon, and Kyung-Soo Kim. Evaluating copy-blend augmentation for low level vision tasks. *arXiv preprint arXiv:2103.05889*, 2021. 1

[64] Pranjay Shyam, Sandeep Singh Sengar, Kuk-Jin Yoon, and Kyung-Soo Kim. Lightweight hdr camera isp for robust perception in dynamic illumination conditions via fourier adversarial networks. *arXiv preprint arXiv:2204.01795*, 2022. 1, 6

[65] Pranjay Shyam, Kuk-Jin Yoon, and Kyung-Soo Kim. Dynamic anchor selection for improving object localization. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9477–9483. IEEE, 2020. 3

[66] Pranjay Shyam, Kuk-Jin Yoon, and Kyung-Soo Kim. Weakly supervised approach for joint object and lane marking detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2885–2895, 2021. 1

[67] Vishwanath A Sindagi, Poojan Oza, Rajeev Yasarla, and Vishal M Patel. Prior-based domain adaptive object detection for hazy and rainy conditions. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 763–780. Springer, 2020. 1

[68] Anu Singha and Mrinal Kanti Bhowmik. Salient features for moving object detection in adverse weather conditions during night time. *IEEE Transactions on circuits and systems for video technology*, 30(10):3317–3331, 2019. 1

[69] Long Sun, Jinshan Pan, and Jinhui Tang. ShuffleMixer: An efficient convnet for image super-resolution. In *Advances in Neural Information Processing Systems*, 2022. 6, 7

[70] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021. 4, 5

[71] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020. 3

[72] Mohsen Vadidar, Ali Kariminezhad, Christian Mayr, Laurent Kloeker, and Lutz Eckstein. Robust environment perception for automated driving: A unified learning pipeline for visual-infrared object detection. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 367–374. IEEE, 2022. 2

[73] Abhinav Valada, Johan Vertens, Ankit Dhall, and Wolfram Burgard. Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4644–4651. IEEE, 2017. 1

[74] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M. Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions, 2021. 1

[75] Johan Vertens, Jannik Zürn, and Wolfram Burgard. Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8461–8468. IEEE, 2020. 2, 6

[76] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-yolov4: Scaling cross stage partial network. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 13029–13038, 2021. 3

[77] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets

new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022. 3

[78] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391, 2020. 3

[79] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 9197–9206, 2019. 3

[80] Kun Wang, Zhenyu Zhang, Zhiqiang Yan, Xiang Li, Baobei Xu, Jun Li, and Jian Yang. Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16055–16064, 2021. 1

[81] Yan Wang. Edge-enhanced feature distillation network for efficient super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 777–785, 2022. 3

[82] Yan Wang. Edge-enhanced feature distillation network for efficient super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 777–785, June 2022. 7

[83] Yi Wang, Syed Muhammad Arsalan Bashir, Mahrukh Khan, Qudrat Ullah, Rui Wang, Yilin Song, Zhe Guo, and Yilong Niu. Remote sensing image super-resolution and object detection: Benchmark and state of the art. *Expert Systems with Applications*, 197:116793, 2022. 3

[84] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2567–2575, 2022. 3

[85] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15769–15778, 2021. 1

[86] Yuxuan Xiao, Aiwen Jiang, Jihua Ye, and Ming-Wen Wang. Making of night vision: Object detection under low-illumination. *IEEE Access*, 8:123075–123086, 2020. 1

[87] Shangliang Xu, Xinxin Wang, Wenyu Lv, Qinyao Chang, Cheng Cui, Kaipeng Deng, Guanzhong Wang, Qingqing Dang, Shengyu Wei, Yuning Du, et al. Pp-yoloe: An evolved version of yolo. *arXiv preprint arXiv:2203.16250*, 2022. 3

[88] Ravi Yadav, Ahmed Samir, Hazem Rashed, Senthil Yogamani, and Rozenn Dahyot. Cnn based color and thermal image fusion for object detection in automated driving. *Irish Machine Vision and Image Processing*, 2020. 2

[89] Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International conference on image processing (ICIP)*, pages 276–280. IEEE, 2020. 2

[90] Jiaqing Zhang, Jie Lei, Weiying Xie, Zhenman Fang, Yunsong Li, and Qian Du. Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. 3

[91] Xingchen Zhang, Ping Ye, Henry Leung, Ke Gong, and Gang Xiao. Object fusion tracking based on visible and infrared images: A comprehensive review. *Information Fusion*, 63:166–187, 2020. 2

[92] Hengyuan Zhao, Xiangtao Kong, Jingwen He, Yu Qiao, and Chao Dong. Efficient image super-resolution using pixel attention. In *European Conference on Computer Vision*, pages 56–72. Springer, 2020. 7

[93] Qijie Zhao, Tao Sheng, Yongtao Wang, Zhi Tang, Ying Chen, Ling Cai, and Haibin Ling. M2det: A single-shot object detector based on multi-level feature pyramid network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9259–9266, 2019. 3

[94] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3

[95] Fuhao Zou, Wei Xiao, Wanting Ji, Kunkun He, Zhixiang Yang, Jingkuan Song, Helen Zhou, and Kai Li. Arbitrary-oriented object detection via dense feature fusion and attention model for remote sensing super-resolution image. *Neural Computing and Applications*, 32:14549–14562, 2020. 3