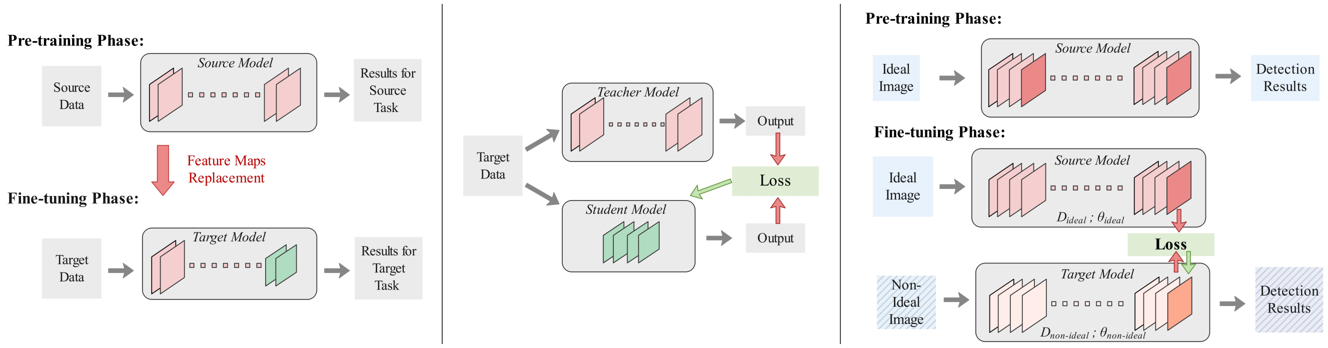
Feature Corrective Transfer Learning: End-to-End Solutions to Object Detection in Non-Ideal Visual Conditions

Chuheng Wei, Guoyuan Wu, Matthew J. Barth
University of California Riverside
chuheng.wei@email.ucr.edu, gywu@cert.ucr.edu, barth@ece.ucr.edu



(a) Transfer Learning

(b) Knowledge Distillation

(c) Feature Correction Transfer Learning

Figure 1. A Visual Comparison of (a) Transfer Learning, (b) Knowledge Distillation, and (c) Feature Correction Transfer Learning

Abstract

A significant challenge in the field of object detection lies in the system's performance under non-ideal imaging conditions, such as rain, fog, low illumination, or raw Bayer images that lack ISP processing. Our study introduces 'Feature Corrective Transfer Learning', a novel approach that leverages transfer learning and a bespoke loss function to facilitate the end-to-end detection of objects in these challenging scenarios without the need to convert non-ideal images into their RGB counterparts. In our methodology, we initially train a comprehensive model on a pristine RGB image dataset. Subsequently, non-ideal images are processed by comparing their feature maps against those from the initial ideal RGB model. This comparison employs the Extended Area Novel Structural Discrepancy Loss (EANSDL), a novel loss function designed to quantify similarities and integrate them into the detection loss. This approach refines the model's ability to perform object detection across varying conditions through direct feature map correction, encapsulating the essence of Feature Corrective Transfer Learning. Experimental validation on variants of the KITTI dataset demonstrates a significant improvement in mean Average Precision (mAP), resulting in a 3.8-8.1% relative enhancement in detection under non-ideal conditions compared to the baseline model, and a less marginal performance difference within 1.3% of the mAP@[0.5:0.95] achieved under ideal conditions by the standard Faster RCNN algorithm.

1. Introduction

As a vital component of computer vision, object detection is used in a wide range of applications such as autonomous driving, surveillance, and augmented reality [35]. Despite significant advancements, the robust detection of objects under non-ideal imaging conditions—such as rain [22], fog [26], low illumination [6], or directly from raw Bayer images [1] without Image Signal Processing (ISP) [28]—remains a considerable challenge. Traditional methods often rely on preprocessing steps to convert non-ideal images into more 'ideal' conditions before detection [19], which can lead to loss of details and introduce unwanted artifacts.

In response to these challenges, transfer learning presents an effective strategy by leveraging pre-existing models trained on extensive, well-labeled datasets to address the variances in imaging conditions [34]. Traditionally, it involves a two-phase process: initially training a source model on comprehensive source data, followed by fine-tuning where parts of this model are adjusted or fixed to adapt to new tasks as shown in Figure 1-(a). While effective in managing image quality variances, existing transfer learning approaches often fall short in addressing the specific challenges posed by non-ideal imaging environments [24].

Knowledge distillation algorithms [12], as shown in Figure 1-(b), primarily utilize a complex model (Teacher) pre-trained on a large dataset to enhance the performance of a smaller model (Student) on a target task. This is achieved by minimizing the loss function based on the discrepancy between their outputs, thus guiding the student model's improvement. Drawing inspiration from both traditional transfer learning and knowledge distillation, we propose the *Feature Corrective Transfer Learning* (FCTL) approach, illustrated in Figure 1-(c). In the pre-trained phase, a comprehensive source model is trained on ideal images. During the fine-tuning phase, the structure and parameters of the source model are kept unchanged while establishing an identical target model. This phase involves training with both non-ideal and ideal versions of the same image, leveraging the established source model to perform feature correction on the target model through a specific loss function at selected layers. This novel approach, FCTL, distinguishes itself by emphasizing direct feature map correction to enhance the robustness and accuracy of object detection models under non-ideal conditions, without necessitating the conversion of non-ideal images to their RGB counterparts.

Based on this framework, we have developed the *Feature Corrective Transfer Learning* NITF-RCNN algorithm that is supplemented by the Extended Region Novel Structure Difference Loss (EANSDL). In our method, we use a two-stage training strategy, establishing a strong baseline using the original RGB dataset and then performing feature map correction on non-ideal image models. By prioritizing direct feature map correction over traditional preprocessing, this process iteratively enhances the model's ability to detect objects under adverse conditions.

1.1. Contributions

- **Feature Corrective Transfer Learning Framework:** A new transfer learning approach is tailored to object detection in challenging conditions, employing feature map correction to align non-ideal images with high-quality RGB datasets, thereby improving robustness and detection accuracy.
- **Non-Ideal Image Transfer Faster-RCNN (NITF-RCNN):** An adaptation of the Faster-RCNN architecture incorporates our feature map correction algorithm, designed to specifically address the challenges presented by non-ideal imaging conditions, ensuring a thoughtful rather than blanket application of transfer learning.
- **Extended Area Novel Structural Discrepancy Loss (EANSDL):** A novel loss function is created to facilitate feature map correction, enabling precise adjustments during training by quantifying the discrepancy between feature maps under different conditions, thus enhancing the model's performance in detecting objects across diverse visual environments.

2. Related Work

2.1. Object Detection under Non-Ideal Visual Conditions

Within the domain of object detection under non-ideal visual conditions, a diverse array of strategies has been explored, ranging from traditional preprocessing to innovative end-to-end models. Sindagi et al. [22] preprocessed images affected by haze and rain using traditional techniques and weather-specific knowledge for object detection. Kvyetnyy et al. [15], alternatively, addressed low-light challenges through denoising methods like bilateral filtering and wavelet thresholding, aiming to improve detection performance. Moreover, some scholars have adopted two-stage model approaches, grounded in deep learning. For example, Huang et al. [13] introduced a dual-subnet network (DSNet), comprising detection and restoration subnets to achieve image restoration and object detection under harsh weather conditions separately. Yang et al. [32] presented a two-stage unsupervised deraining approach, utilizing non-local contrastive learning to decouple the rain layer from clean images more effectively before object detection tasks.

Emerging research, however, aims to develop truly end-to-end solutions. For example, Wang et al. [27] proposed an end-to-end object detection network to mitigate the impact of rainfall, featuring cascaded networks for image restoration and object detection. While this is an end-to-end training approach, deraining and detection are still separated into two networks, with the first network producing derained images. Additionally, Wei et al. [29] attempted to perform end-to-end object detection on raw images by incorporating camera parameters into the network to adapt to the features of raw Bayer images. However, This method requires other forms of input besides images and more complex neural networks.

In summary, there is a conspicuous absence of suitable, truly end-to-end models that forego the intermediate image restoration step. There also lacks a universal model capable of effectively handling all types of non-ideal conditions, highlighting a significant gap in the current state of object detection under non-ideal visual conditions.

2.2. Advancing Object Detection through Transfer Learning Techniques

In the field of computer vision, transfer learning has emerged as a key strategic tool for improving performance in related, yet distinct, tasks [24]. To enhance model performance and efficiency, researchers have begun applying the principles of transfer learning to object detection since its introduction. To improve the accuracy of object detection, Ito et al. [14] used genetic algorithms within the transfer learning process to determine which layers should be relearned automatically. They also avoided the trial-and-error approach inherent in traditional approaches to object detection. Their research demonstrates that partially intercepting neural networks can enhance the efficiency of transfer learning

Several studies have focused on transferring abilities learned from large, general datasets to more specialized tasks, such as detecting small objects. A resolution adaptation scheme was employed by Xu et al. [31] to enhance the detection of small-scale objects by adjusting models trained on generic datasets using images of various smaller resolutions, thereby significantly increasing performance. According to Bu et al. [2], a transfer learning system named *GAIA* was designed in recognition of the unique requirements of the object detection domain. Using this system, tailored solutions are automatically generated based on heterogeneous downstream demands, offering powerful pretrained weights and selecting models that meet specific needs for object detection, including latency constraints and particular data domains, thereby demonstrating a significant advancement in small-sample object detection algorithms.

Additionally, efforts have been made to extend the utility of transfer learning beyond mere knowledge transfer by utilizing data augmentation and synthetic datasets. Talukdar et al. [24] explored the generation of synthetic datasets through various data augmentation algorithms to assist in the transfer learning process for convolutional neural networks. Their experiments across a range of object detection algorithms validated the significance of synthetic datasets in enhancing transfer learning outcomes. Moreover, some researchers have applied transfer learning to address challenges presented by non-ideal data conditions. For instance, Chen et al. [5] improved the Faster R-CNN algorithm through transfer learning, employing two domain adaptation structures to measure domain similarity. Their *Domain Adaptation Faster R-CNN* which utilized adversarial training, proved effective in low-light conditions.

However, these approaches primarily focus on domain-specific challenges and do not extensively explore object detection under adverse weather conditions through transfer learning. This gap highlights the novelty of the proposed framework, which specifically addresses feature map correction for object detection in challenging environmental conditions, setting a new research direction in this area.

3. Feature Corrective Transfer Learning Framework

End-to-end object detection holds paramount importance in computer vision, offering a seamless process from image input to object identification and localization [3]. Traditionally, object detection under non-ideal conditions such as poor lighting, adverse weather, or unprocessed raw images, necessitates transforming these images into a more ’ideal’ state before detection can occur. However, this transformation often targets human visual preferences rather than the requirements of neural networks. This paper posits that true end-to-end detection should circumvent the need for such transformations, thereby directly addressing the detection in non-ideal conditions.

However, evidence suggests that models developed for ideal situations do not perform optimally on non-ideal cases [4], which underscores the necessity for model adjustments [29]. Direct modifications to handle non-ideal image features could introduce bias or overfitting to specific conditions. To address this, we propose *Feature Map Correction* (FMC) to assist the neural network training process without altering the underlying architecture. Most object detection algorithms process the input image through a neural network to perform bounding box regression and object classification across various feature layers and scales. The *Feature Corrective Transfer Learning* (FCTL) method introduced in this paper aims to guide the training of models on non-ideal images towards closer alignment with the feature layers of models trained on ideal images.

3.1. Implementation of the FCTL Framework

To formalize the FCTL framework, we define a mathematical model consisting of the following key steps:

1. **Model Selection and Training on Ideal Images**
   Let D_{ideal} represent the dataset of ideal images. We select an object detection model M, and train it on D_{ideal} to optimize the model parameters θ_{ideal}:

   $$\theta_{\text{ideal}} = \arg\min_{\theta} \mathcal{L}_{\text{det}}(M(D_{\text{ideal}}; \theta)) \tag{1}$$

   where $\mathcal{L}_{\text{det}}$ is the total loss function for the object detection task, typically comprising classification loss and bounding box regression loss.
2. **Generation of Non-Ideal Image Versions**
   For each ideal image $x \in D_{\text{ideal}}$, we generate a non-ideal version x' by synthesizing non-ideal conditions such as rainy weather. This can be achieved by adding noises, blurring, etc.
3. **Training the Same Object Detection Model on Non-Ideal Images**

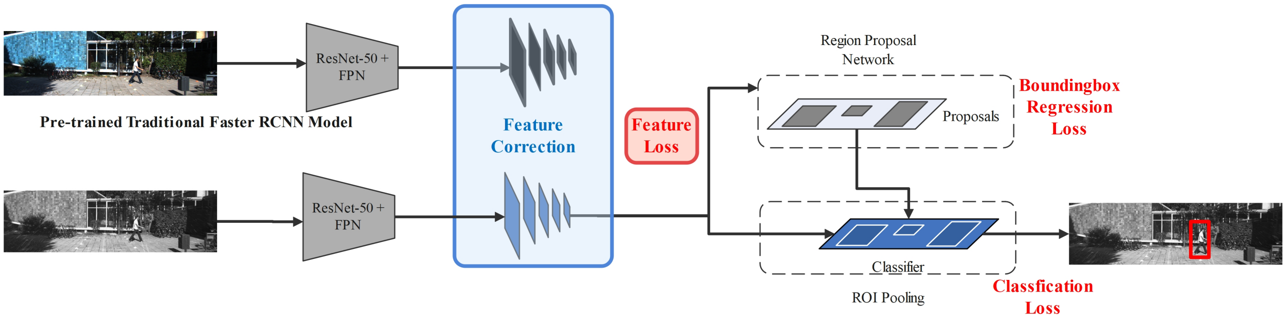


Figure 2. Architecture of Non-Ideal Image Transfer Faster RCNN (NITF-RCNN) Model

Next, the same model M is trained on the non-ideal images $D_{\text{non\_ideal}}$, while also using the corresponding ideal images for validation. During this phase, one or multiple feature layers are selected to assess the similarity between the feature maps of the model trained on ideal images and those of the model being trained on non-ideal images, employing a feature similarity loss function $\mathcal{L}_{\text{fs}}$:

$$\theta_{\text{non\_ideal}} = \arg\min_{\theta} \Bigg(\mathcal{L}_{\text{det}}(M(D_{\text{non\_ideal}}; \theta)) \\ + \lambda \mathcal{L}_{\text{fs}}(F_{\text{ideal}}, F_{\text{non\_ideal}}) \Bigg). \quad (2)$$

Here, F_{ideal} and $F_{\text{non\_ideal}}$ represent the feature maps from the model under ideal and non-ideal conditions, respectively, and λ is a coefficient that balances the two loss terms.

4. **Incorporating Feature Similarity Loss during Backpropagation**
During the backpropagation process in training, the total loss $\mathcal{L}_{\text{total}}$ to be minimized includes not only the standard detection loss $\mathcal{L}_{\text{det}}$—comprising classification loss and bounding box regression loss—but also the feature similarity loss $\mathcal{L}_{\text{fs}}$. This dual-objective loss function aims to ensure accuracy in object detection while introducing a mechanism for feature map correction:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \lambda \mathcal{L}_{\text{fs}}, \quad (3)$$

$$\theta = \arg\min_{\theta} \mathcal{L}_{\text{total}}. \quad (4)$$

The feature similarity loss $\mathcal{L}_{\text{fs}}$ is designed to effectively measure the discrepancies in structure and content between the feature maps of the model trained on ideal images and those trained on non-ideal images. It is crucial to note that similarity in feature space can significantly differ from image similarity, necessitating a distinct evaluation metric. This paper introduces the *Extended Area Novel Structural Discrepancy Loss* (EANSDL) method to assess the similarity at the feature level.

In subsequent section, we elaborate on the modifications to the Faster RCNN framework, leading to the development of the Non-Ideal Image Transfer Faster-RCNN (NITF-RCNN) model. This model incorporates a feature similarity loss to evaluate the similarity of pyramid feature maps, showcasing the practical application of the FCTL methodology.

4. Methodology

4.1. Non-Ideal Image Transfer Faster R-CNN (NITF-RCNN)

The NITF-RCNN framework adapts the traditional Faster R-CNN [8] to object detection in non-ideal visual conditions, maintaining the original architecture while incorporating feature map correction. As shown in Figure 2, this specialized algorithm employs a dual backbone structure: a static backbone derived from a model pre-trained on ideal images and a dynamic backbone that is fine-tuned on non-ideal images.

Training on Ideal Images

The foundation is laid by training a Faster R-CNN model equipped with a ResNet-50 backbone [11] and Feature Pyramid Network (FPN) [17] on a dataset of ideal RGB images to establish the static backbone. The objective function for this training phase is defined as:

$$\theta_{\text{ideal}} = \arg\min_{\theta} \mathcal{L}_{\text{Faster R-CNN}}(D_{\text{ideal}}; \theta). \quad (5)$$

Feature Corrective Transfer Learning

In the feature-level transfer learning phase, the pre-trained static backbone extracts feature maps from the ideal images to form the ”ideal pyramid,” while the non-ideal images are concurrently processed through the dynamic backbone. Both are then subject to the region proposal network (RPN) [8] and Region of Interest (ROI) Pooling [9]. The dynamic backbone undergoes training, guided by the following combined loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}}(D_{\text{non\_ideal}}; \theta) + \lambda \mathcal{L}_{\text{EANSDL}}(F_{\text{ideal}}, F_{\text{non\_ideal}}). \quad (6)$$

where $\mathcal{L}_{\text{det}}$ represents the standard object detection loss, which includes bounding box regression and classification losses [8]. $\mathcal{L}_{\text{EANSDL}}$ denotes the *Extended Area Novel Structural Discrepancy Loss*, a dedicated loss function assessing the similarity between feature maps. λ serves as the balancing coefficient for the feature similarity loss.

Validation

During validation, only the non-ideal images are processed through the NITF-RCNN to assess the model's detection capability in non-ideal conditions. This is achieved by applying the trained model to non-ideal images, with the loss functions serving as indicators of performance:

$$\mathcal{L}_{\text{validation}}(D_{\text{non\_ideal}};\theta). \tag{7}$$

Structural Summary and Potential Advantages

The NITF-RCNN framework retains the integrity of the traditional Faster R-CNN structure, with the addition of the feature correction component for ideal images. This approach provides several potential advantages:

- **Feature Correction Without Structural Modification:** The framework enhances object detection under non-ideal conditions without the need for significant modifications to the existing architecture.
- **Direct Feature-Level Adaptation:** By correcting feature maps directly through the FCTL approach, the model is better equipped to handle environmental disturbances inherent in non-ideal images.
- **Balanced Learning:** The use of a joint loss function allows the model to balance feature correction with the primary detection tasks, potentially improving generalization and robustness.

4.2. Adaptive Structural Alignment via EANSDL

To address the challenge of aligning and comparing feature maps under complex, non-ideal conditions, we introduce *Extended Area Novel Structural Discrepancy Loss* (EANSDL) which not only identifies pixel-level discrepancies but also ensures the structural integrity across larger areas, making it beneficial for advanced object detection frameworks like Faster RCNN in less-than-ideal conditions. EANSDL conducts a comprehensive evaluation, rectifying immediate discrepancies while maintaining overall structural coherence, significantly enhancing Faster RCNN's detection precision. Its design adaptively balances the analysis between detailed discrepancies and broader alignments, dynamically adjusting the gradient consistency evaluation across the feature pyramid's hierarchical layers. This method achieves heightened sensitivity to layer-specific scales and resolutions, thereby bolstering structural integrity and ensuring robust object detection across diverse imaging conditions.

4.2.1 Mathematical Formulation

Consider two feature maps, A (non-ideal conditions) and B (ideal conditions), each with dimensions $[batchsize, channels, width, height]$. The formulation of EANSDL is given by:

$$\begin{aligned}&\text{EANSDL}(A,B,\delta,r_{\mathcal{L}})\\&=D(\delta)\cdot\frac{1}{W\cdot H}\sum_{x=1}^{W}\sum_{y=1}^{H}\Bigg(\exp(-\Delta S(x,y))\cdot\Delta S(x,y)\\&+\lambda\cdot\Omega(A,B,x,y,r_{\mathcal{L}})\Bigg).\end{aligned}\tag{8}$$

where:

- $D(\delta)$ introduces a time-varying attenuation factor that adjusts the sensitivity of the loss function to training progress,with τ denoting the ratio of the current epoch to total epochs (ensuring that the loss adapts throughout the training lifecycle).
- $\Delta S(x,y)$ denotes the local gradient magnitude difference at position (x,y), capturing immediate structural variances.
- λ represents a balancing factor for the contribution of the extended area gradient consistency.
- $\Omega(A,B,x,y,r_{\mathcal{L}})$ encapsulates the extended area gradient consistency across a neighborhood radius $r_{\mathcal{L}}$, dynamically adjusted for each level of the Faster RCNN feature pyramid as:

$$r_{\mathcal{L}}=r_0/2^{level}, \tag{9}$$

where r_0 is the initial radius at the largest feature map, and $level$ denotes the specific layer within the feature pyramid.

4.2.2 Implementation Details

Time-varying Attenuation Factor

The *Time-varying Attenuation Factor*, represented as $D(\delta)$, introduces a dynamic mechanism to adjust the responsiveness of the loss function throughout the training duration. The term δ signifies the proportion of the current epoch relative to the total number of epochs, calculated as:

$$\delta=\frac{\text{current\_epoch}}{\text{total\_epochs}}. \tag{10}$$

The implementation of this factor facilitates a methodological shift in the model's emphasis from rectifying prominent structural disparities in the initial training phase to honing finer details in subsequent phases of the training process.

$D(\delta)$ is delineated as follows:

$$D(\delta) = \exp(-\alpha \cdot \delta^{\beta}), \tag{11}$$

where:

- α regulates the initial steepness of the decay trajectory;
- β adjusts the curvature to slow down the decay pace.

Fundamentally, $D(\delta)$ empowers the model to initially concentrate on correcting significant mismatches between feature maps, ensuring a solid foundation is established. As training progresses and the model evolves in complexity, the attenuation factor reduces the emphasis on these mismatches. This modification aids in diminishing the influence of the EANSDL on the total loss for object detection during the later phases, allowing for a greater focus on the quintessential tasks of object detection.

Gradient Computation Function

Central to EANSDL, the gradient computation function $G(\cdot)$ employs the Sobel operator [23] to delineate edges and structural attributes across the feature maps. This operator convolves the feature map with two distinct 3×3 kernels, each engineered to unearth edges along respective orientations:

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad S_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}. \tag{12}$$

Vertical edges are identified through the horizontal gradient (Sobel-x), whereas horizontal edges are pinpointed by the vertical gradient (Sobel-y). The formulas for calculating these gradients are as follows:

$$G_x(A) = A * S_x, \quad G_y(A) = A * S_y. \tag{13}$$

The aggregate gradient magnitude is consequently determined by amalgamating these orthogonal gradients:

$$G(A, x, y) = \sqrt{G_x(A, x, y)^2 + G_y(A, x, y)^2}. \tag{14}$$

Local Gradient Magnitude Difference

The local gradient magnitude difference between feature maps A and B, represented by $/DeltaS(x, y)$, is expressed as:

$$\Delta S(x, y) = |G(A, x, y) - G(B, x, y)|. \tag{15}$$

This metric quantifies the direct structural disparities, highlighting areas where edge and texture information significantly differ due to non-ideal imaging conditions. Essentially, $\Delta S(x, y)$ pinpoints the local discrepancies that the model needs to correct to better align feature maps derived from non-ideal and ideal scenarios. The term $\exp(-\Delta S(x, y))$ acts as a weighting factor, modulating the contribution of each local discrepancy $\Delta S(x, y)$ to the overall loss.

When these two terms are multiplied, i.e., $\exp(-\Delta S(x, y)) \cdot \Delta S(x, y)$, the exponential decay function in the loss calculation magnifies the impact of smaller discrepancies, directing the model's focus on refining minor but essential structural differences. Simultaneously, it lessens the penalty on larger discrepancies to avoid undue penalization for less critical variances. This mechanism ensures a balanced model training, prioritizing major discrepancies in early phases for overall performance and shifting towards finer adjustments in later stages as feature map discrepancies diminish, facilitating nuanced structural alignment for improved object detection accuracy.

Extended Area Gradient Consistency

The Extended Area Gradient Consistency term, $\Omega(A, B, x, y, r_{\mathcal{L}})$, scrutinizes the uniformity of gradient transitions within a specified vicinity, thereby assessing broader spatial patterns. It evaluates the consistency of gradient changes across an extended neighborhood, defined by a radius $r_{\mathcal{L}}$. This radius is adaptively adjusted for each layer in the Faster RCNN feature pyramid, allowing for a multi-scale analysis:

$$\begin{aligned} \Omega(A, B, x, y, r_{\mathcal{L}}) &= \frac{1}{(2r_{\mathcal{L}} + 1)^2} \\ &\cdot \sum_{i=-r_{\mathcal{L}}}^{r_{\mathcal{L}}} \sum_{j=-r_{\mathcal{L}}}^{r_{\mathcal{L}}} |(G(A, x, y) - G(A, x+i, y+j)) \\ &-(G(B, x, y) - G(B, x+i, y+j))|. \end{aligned} \tag{16}$$

This extended area gradient consistency ensures that the model not only captures pixel-by-pixel discrepancies, but also appreciates broader spatial patterns and alignments. This multi-scale approach is critical for robust object detection, as it allows the model to recognize and adapt to the variances in object sizes and shapes across different feature map scales.

In summary, EANSDL represents a significant advance in object detection, offering powerful structural insight and correction capabilities. By skillfully combining evaluations of both immediate and broader spatial contexts, the EANSDL function empowers object detection algorithms that correct feature maps layers through Transfer Learning, such as NITF RCNN, to deliver unparalleled performance. This approach ensures the meticulous alignment and refinement of feature maps, transcending the challenges posed by non-ideal imaging conditions. This approach not only enhances the model's detection capabilities but also sets a benchmark for feature map analysis and correction in complex visual environments.

Dataset	Faster RCNN			NITF RCNN		
	mAP@0.5	mAP@0.75	mAP@[0.5,0.95]	mAP@0.5	mAP@0.75	mAP@[0.5,0.95]
KITTI	80.87%	61.45%	55.43%	—-	—-	—-
Rainy-KITTI	73.17%	56.15%	49.28%	79.16%(↑8.1%)	59.35%(↑5.6%)	51.86%(↑5.2%)
Foggy-KITTI	71.88%	48.74%	42.31%	75.01%(↑4.4%)	51.30%(↑5.5%)	44.48%(↑5.1%)
Dark-KITTI	62.74%	41.39%	37.89%	65.61%(↑4.6%)	44.05%(↑6.4%)	40.07%(↑5.7%)
Raw-KITTI	76.13%	58.50%	51.76%	79.06%(↑3.8%)	61.91%(↑5.8%)	54.13%(↑4.6%)

Table 1. Evaluation Results of Faster RCNN and NITF RCNN on Different Datasets (The relative percentage improvement on each metric is compared between the NITF RCNN algorithm and Faster RCNN.)

5. Experiments

5.1. Dataset Selection and Generation

In our experimental setup, we combine original and synthesized datasets due to the requirement of having both ideal and non-ideal versions of the same image for transfer learning. This necessity arises from the need to train on ideal images and then adapt to non-ideal conditions. Real non-ideal images, altered to simulate 'ideal' conditions, often lose crucial details due to inherent visual obstructions like rain or fog. Therefore, to maintain content consistency and to ensure a robust training foundation, we opt for real-world images as our ideal dataset and generate synthetic counterparts for the non-ideal scenarios, effectively using high-quality originals to produce less information-dense but contextually aligned images.

The KITTI 2D object detection dataset [7] serves as the foundation for our experiments, known for its real-world driving scenarios, diverse object annotations, and complex urban environments. As the ideal dataset, we utilize KITTI alongside four synthetic datasets—Rainy-KITTI, Foggy-KITTI, Dark-KITTI, and RAW-KITTI—as our non-ideal datasets.

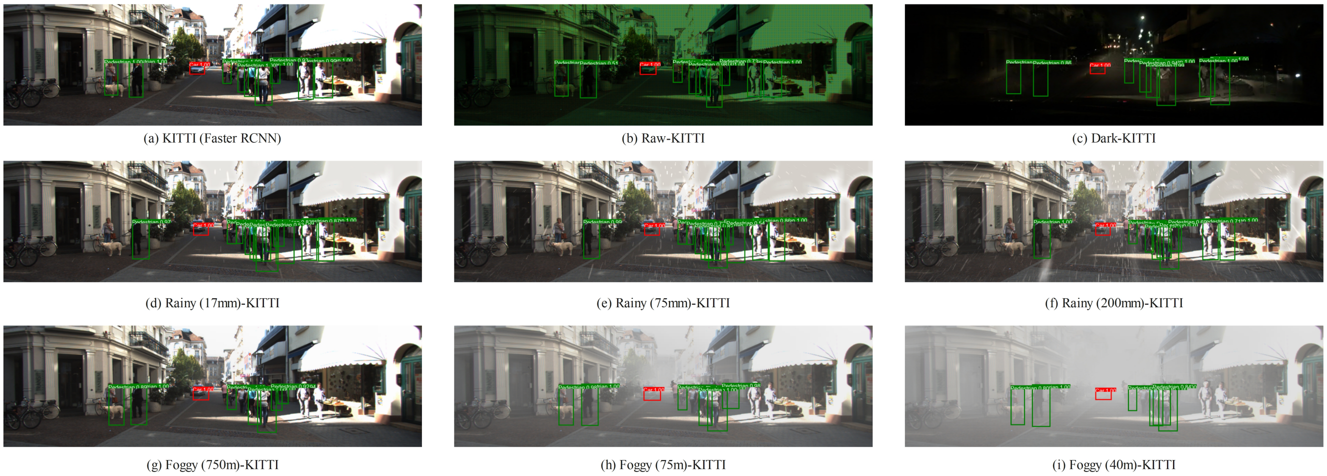
- **Rainy-KITTI & Foggy-KITTI:** For simulating rain and fog conditions, we selected the Rainy-KITTI and Foggy-KITTI datasets [10, 25], recognized for their realistic emulation of these weather effects. The Rainy-KITTI dataset encompasses images under seven distinct rain intensities, ranging from light to heavy downpours. Similarly, the Foggy-KITTI dataset includes images under seven different visibility conditions due to fog. For our experiments, we randomly select an image from each of these conditions to compile our dataset.
- **Dark-KITTI:** To generate a dataset simulating low-light conditions, we followed Rashed et al.'s methodology [21], utilizing the UNIT [18] algorithm for its superior performance in creating realistic night-time images. Using 2000 clear-day images from the KITTI dataset [7] and 2000 night images from the BDD100K dataset [33], we trained a day-to-night model on UNIT and generated the Dark-KITTI dataset.
- **Raw-KITTI:** Addressing the challenge of replicating RAW Bayer images, due to the irreversible nature of the Image Signal Processing (ISP) [28], we adopted a dataset generation method from [4] to create a synthetic color Bayer image dataset, termed Raw-KITTI. This dataset features color channels in the RGB format, ensuring consistency in channel count across all datasets used in our experiments by assigning corresponding colors to each channel of the RAW data.

5.2. Quantitative Results

In our experiment, the learning rate was set to 0.005 and the batch size was configured at 8. We allocated 80% of each dataset for training and reserved 20% for validation, conducting the training over 100 epochs. The performance evaluation was based on the mean Average Precision (mAP), in accordance with the COCO detection benchmark standards [16]. Our model's performance was evaluated using three key metrics: mAP@0.5, mAP@0.75, and mAP@[0.5:0.95]. The mAP@0.5 and mAP@0.75 metrics represent the mean average precision at Intersection over Union (IoU) thresholds of 0.5 and 0.75, respectively, demanding closer alignment with ground truth for higher values. The mAP@[0.5:0.95] metric, averaging performance across an IoU threshold range from 0.5 to 0.95 with 0.05 increments, provides a comprehensive assessment of model accuracy at various levels of detection precision.

Table 1 encapsulates the evaluation results of Faster RCNN and our NITF-RCNN model across different datasets. It highlights the comparative performance improvements of NITF-RCNN over Faster RCNN under various conditions, showcasing the effectiveness of our FCTL framework.

The results clearly demonstrate the superior performance of NITF-RCNN across all non-ideal imaging conditions, with notable performance gains. Specifically, NITF-RCNN achieved an 8.1% increase in mAP@0.5, a 5.6% increase in



(a) KITTI (Faster RCNN)

(b) Raw-KITTI

(c) Dark-KITTI

(d) Rainy (17mm)-KITTI

(e) Rainy (75mm)-KITTI

(f) Rainy (200mm)-KITTI

(g) Foggy (750m)-KITTI

(h) Foggy (75m)-KITTI

(i) Foggy (40m)-KITTI

Figure 3. Detection Results of NITF-RCNN on Derivative Images of ID 000332 from the KITTI Dataset, where (a) represents the original image from the KITTI dataset detected using the Faster RCNN algorithm for comparison.

mAP@0.75, and a 5.2% improvement in mAP@[0.5:0.95] on the Rainy-KITTI dataset. Similar improvements are observed across Foggy-KITTI, Dark-KITTI, and Raw-KITTI datasets, underlining the model's enhanced detection capabilities in challenging visual scenarios.

It is noteworthy that across all evaluated datasets, NITF-RCNN exhibits consistent improvements over Faster RCNN in the comprehensive mAP@[0.5:0.95] metric, with relative gains ranging from 4.6% to 5.7%. This underscores the effectiveness of the NITF-RCNN model in maintaining high accuracy across various levels of detection precision, especially in non-ideal imaging conditions. Remarkably, on the Raw-KITTI dataset, the mAP@[0.5:0.95] performance of the NITF-RCNN approaches that of the ideal KITTI dataset on the Faster RCNN, with a mere 1.3% difference. This highlights the significant advancements made by NITF-RCNN in closing the gap between object detection performances in ideal versus non-ideal conditions, showcasing its potential to operate effectively across a broader range of real-world scenarios.

5.3. Qualitative Results

Figure 3 displays the detection outcomes on four derivative datasets from the KITTI dataset, specifically for image ID 000332. For the Rainy-KITTI and Foggy-KITTI datasets, we showcase detection results across three different levels of rainfall intensity and varying visibility, respectively.

As a result of integrating the insights derived from Figure 3 and Table 1, we are able to demonstrate that our methodology yields performance similar to that of Faster RCNN under ideal conditions for the Rainy KITTI and Raw KITTI datasets. In contrast, the performance on the Dark KITTI and Foggy KITTI datasets is relatively inferior. It is hypothesized that the main reason for this discrepancy is that Rainy and Raw KITTI images are more visual discernible than low-light and foggy images, which facilitates easier detection [20, 30].

6. Conclusion

This research introduces a pioneering approach in computer vision, particularly in object detection under non-ideal conditions such as low light, adverse weather, or directly from raw Bayer images without ISP. Using the novel concept of FCTL in conjunction with a unique loss function, EANSDL, we demonstrate that the NITF-RCNN model is able to significantly improve the object detection ability in various challenging environments. Our methodology bypasses traditional preprocessing requirements for non-ideal images, directly refining the model's feature maps to closely align with those obtained from pristine RGB datasets. Experimental results demonstrate the efficacy of this approach, which shows a substantial improvement in mAP over conventional methods, thus setting an industry benchmark.

This study not only strengthens the robustness and accuracy of object detection models under diverse environmental conditions, but also provides new avenues for further research. It is possible for FCTL to be applied outside of autonomous driving, surveillance, and augmented reality, suggesting its potential for other areas where visual data is compromised by conditions that are not ideal. In future research, FCTL may be adapted to address a wider range of imaging challenges, loss function optimization for higher efficiency, and integration of this approach with other frameworks for object detection. The successful application of FCTL heralds a paradigm shift in how visual data is processed, and promises advancements in various applications reliant on accurate and reliable object detection.

References

[1] Bryce E Bayer. Color imaging array, 1976. US Patent 3,971,065. 1

[2] Xingyuan Bu, Junran Peng, Junjie Yan, T. Tan, and Zhaoxiang Zhang. Gaia: A transfer learning system of object detection that fits your needs. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 274–283, 2021. 3

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[4] Pak Hung Chan, Chuheng Wei, Anthony Huggett, and Valentina Donzella. Raw camera data object detectors: an optimisation for automotive processing and transmission. *Authorea Preprints*, 2023. 3, 7

[5] Jinyong Chen, Jianguo Sun, Yuqian Li, and Changbo Hou. Object detection in remote sensing images based on deep transfer learning. *Multimedia Tools and Applications*, 81: 12093 – 12109, 2021. 3

[6] Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask aet with orthogonal tangent regularity for dark object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2553–2562, 2021. 1

[7] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 7

[8] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 4, 5

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 4

[10] Shirsendu Sukanta Halder, Jean-François Lalonde, and Raoul de Charette. Physics-based rendering for improving robustness to rain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10203–10212, 2019. 7

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[13] Shih-Chia Huang, Trung-Hieu Le, and Da-Wei Jaw. Dsnet: Joint semantic learning for object detection in inclement weather conditions. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2623–2633, 2020. 2

[14] Ryuji Ito, H. Nobuhara, and S. Kato. Transfer learning method for object detection model using genetic algorithm. *J. Adv. Comput. Intell. Intell. Informatics*, 26:776–783, 2022. 3

[15] Roman Kvyetnyy, Roman Maslii, Volodymyr Harmash, Ilona Bogach, Andrzej Kotyra, Żaklin Gradz, Aizhan Zhanpeisova, and Nursanat Askarova. Object detection in images with low light condition. In *Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2017*, pages 250–259. SPIE, 2017. 2

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 7

[17] 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. 4

[18] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017. 7

[19] 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*, pages 1792–1800, 2022. 1

[20] Nguyen Anh Minh Mai, Pierre Duthon, Louahdi Khoudour, Alain Crouzil, and Sergio A Velastin. 3d object detection with sls-fusion network in foggy weather conditions. *Sensors*, 21(20):6711, 2021. 8

[21] Hazem Rashed, Mohamed Ramzy, Victor Vaquero, Ahmad El Sallab, Ganesh Sistu, and Senthil Yogamani. Fusemodnet: Real-time camera and lidar based moving object detection for robust low-light autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 7

[22] 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, 2

[23] Irwin Sobel, Gary Feldman, et al. A 3x3 isotropic gradient operator for image processing. *a talk at the Stanford Artificial Project in*, 1968:271–272, 1968. 6

[24] Jonti Talukdar, Sanchit Gupta, PS Rajpura, and Ravi S Hegde. Transfer learning for object detection using state-of-the-art deep neural networks. In *2018 5th international conference on signal processing and integrated networks (SPIN)*, pages 78–83. IEEE, 2018. 2, 3

[25] Maxime Tremblay, Shirsendu Sukanta Halder, Raoul De Charette, and Jean-François Lalonde. Rain rendering for evaluating and improving robustness to bad weather. *International Journal of Computer Vision*, 129:341–360, 2021. 7

[26] Chengjia Wang, Shizhou Dong, Xiaofeng Zhao, Giorgos Papanastasiou, Heye Zhang, and Guang Yang. Saliencygan: Deep learning semisupervised salient object detection in the fog of iot. *IEEE Transactions on Industrial Informatics*, 16 (4):2667–2676, 2019. 1

[27] Kaige Wang, Tianming Wang, Jianchuang Qu, Huatao Jiang, Qing Li, and Lin Chang. An end-to-end cascaded image deraining and object detection neural network. *IEEE Robotics and Automation Letters*, 7(4):9541–9548, 2022. 2

[28] Chuheng Wei. Vehicle detecting and tracking application based on yolov5 and deepsort for bayer data. In *2022 17th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pages 843–849. IEEE, 2022. 1, 7

[29] Chuheng Wei, Guoyuan Wu, Matthew Barth, Pak Hung Chan, Valentina Donzella, and Anthony Huggett. Enhanced object detection by integrating camera parameters into raw image-based faster r-cnn. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 4473–4478. IEEE, 2023. 2, 3

[30] 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. 8

[31] Xinkai Xu, Hailan Zhang, Yan Ma, Kang Liu, Hong Bao, and Xu Qian. Transdet: Toward effective transfer learning for small-object detection. *Remote Sensing*, 2023. 3

[32] Mingxuan Yang, Xiyu Han, Xianyao Ping, Zipeng Li, and Jing Xiao. A clearer image: Improving object detection in real rainy conditions with two-stage processing. *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 57–62, 2023. 2

[33] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 7

[34] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 1

[35] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023. 1