HalluciDet: Hallucinating RGB Modality for Person Detection Through Privileged Information

Heitor Rapela Medeiros, Fidel A. Guerrero Peña, Masih Aminbeidokhti
Thomas Dubail, Eric Granger, Marco Pedersoli
LIVIA, Dept. of Systems Engineering, ETS Montreal, Canada
{heitor.rapela-medeiros.1, fidel-alejandro.guerrero-pena}@ens.etsmtl.ca
{masih.aminbeidokhti.1, thomas.dubail.1}@ens.etsmtl.ca
{eric.granger, marco.pedersoli}@etsmtl.ca

Abstract

A powerful way to adapt a visual recognition model to a new domain is through image translation. However, common image translation approaches only focus on generating data from the same distribution as the target domain. Given a cross-modal application, such as pedestrian detection from aerial images, with a considerable shift in data distribution between infrared (IR) to visible (RGB) images, a translation focused on generation might lead to poor performance as the loss focuses on irrelevant details for the task. In this paper, we propose HalluciDet, an IR-RGB image translation model for object detection. Instead of focusing on reconstructing the original image on the IR modality, it seeks to reduce the detection loss of an RGB detector, and therefore avoids the need to access RGB data. This model produces a new image representation that enhances objects of interest in the scene and greatly improves detection performance. We empirically compare our approach against state-of-the-art methods for image translation and for fine-tuning on IR, and show that our HalluciDet improves detection accuracy in most cases by exploiting the privileged information encoded in a pre-trained RGB detector. Code: `https://github.com/heitorrapela/HalluciDet`.

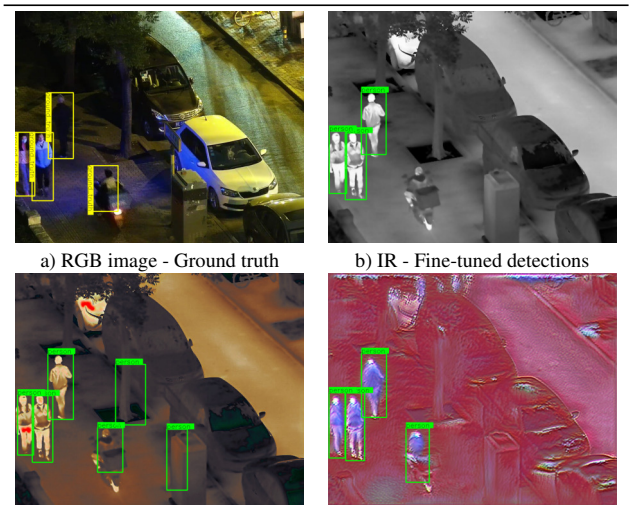


Figure 1. Example of detections using baseline and HalluciDet methods on LLVIP data. (a) Original RGB image with ground truth annotations (yellow). (b) IR image with corresponding detections of a fine-tuned model (green). (c) Translated image from IR to RGB produced by FastCUT and corresponding RGB detections (green). (d) Hallucinated image produced by our method and RGB detections (green); HalluciDet does not seek to reconstruct all image details but only to enhance the objects of interest.

1. Introduction

The proliferation of hardware sensors has greatly advanced the collection of large-scale datasets. Such datasets have significantly improved the performance of deep learning (DL) algorithms across various fields, including surveillance [2], industrial monitoring [16], self-driving cars [31], and robotics [25]. By providing high-resolution data, these sensors offer additional observations of common environmental phenomena to aid in the effectiveness of DL algorithms [26].

The additional information from different sensors has been employed in diverse settings [13, 30]. In computer vision applications, combining sensors with distinct environmental sensing perspectives, such as varying points of view and modality sensing information, can increase model performance, enabling possibilities that were previously unavailable. Furthermore, in the context of self-driving cars and intelligent building applications, two modalities commonly used are visible (RGB) and infrared (IR) [32]. In par-

ticular, the RGB modality offers valuable information for tasks like object detection, which generates bounding boxes for target objects within colored images. These colored images are known to have more diverse information due to their characteristics on the RGB light spectrum, especially in the presence of light. Thus, these RGB sensors are preferred to be used in daily activities where there is the presence of sunlight. On the other hand, the IR spectrum provides additional information for the visible modality when the light is low, especially during the night [15], and also complementary information, primarily related to thermal sensing. Furthermore, IR is vastly applied in surveillance applications [38], which require the device to capture information in light-restricted environments. IR object detection is known to detect objects using IR radiation emitted from the object, which varies depending on the object's material.

Despite the impressive performance of DL models, their effectiveness can significantly deteriorate when applied to modalities that were not present during the training [3, 33]. For example, a model trained on RGB images may not perform well on IR images during testing [35]. To address the issue, some studies utilize image-to-image translation techniques to narrow the gap between modalities distributions. Typically, these methods employ classical pixel manipulation techniques or deep neural networks to generate intermediate representations, which are then fed into a detector trained on the source modality. However, transitioning from IR to RGB has proven challenging due to generating color information while filtering out non-meaningful data associated with diverse heat sources. This challenge is particularly pronounced when the target category is also a heat-emitting source, such as a person.

In this work, we argue that achieving a robust intermediate representation for a given task needs guiding the image-to-image translation using a task-specific loss function. Here, we introduce HalluciDet, a novel approach for image translation focusing on detection tasks. Inspired by the learning using privileged information (LUPI) paradigm [34], we utilize a robust people detection network previously trained on an RGB dataset to guide our translation process from IR to RGB. Our translation approach relies on an annotated IR dataset and an RGB detector to identify the appropriate representation space. The ultimate goal is to find a translation model, hereafter referred to as the Hallucination network, capable of translating IR images into meaningful representation to achieve accurate detections with an RGB detector.

Our main contributions can be summarized as follows:

(1) We propose HalluciDet, a novel approach that leverages privileged information from pre-trained detectors in the RGB modality to guide end-to-end image-to-image translation for the IR modality.

(2) Given that our model focuses on the IR detection task, HalluciDet uses a straightforward yet powerful image translation network to reduce the domain gap between IR-RGB modalities, guided by the proposed hallucination loss function incorporating standard object detection terms.

(3) Through experiments conducted on two challenging IR-RGB datasets (LLVIP and FLIR ADAS), we compare HalluciDet against various image-to-image translation and traditional pixel manipulation methods. Our approach is seen to improve detection accuracy on the IR modality by incorporating privileged information from RGB.

2. Related Work

Object detection. Different from classification tasks, in which we want only to classify the object category, in object detection, additionally, the task is to know specific positions of the objects [36]. Deep learning object detection methods are categorized as two-stage and one-stage detectors. The two-stage detector extracts regions of interest or proposals for a second-stage classifier. Then, the second stage is responsible for classifying if there is an object in that region. One commonly used two-stage detector is the Faster R-CNN proposed by [27]. It is the first end-to-end DL object detector to reach real-time speed. The speedup was achieved by introducing the Region Proposal Network (RPN), a network responsible for the region proposals without impacting the computational performance compared with previous region proposals algorithms [28]. The one-stage detectors mainly focus on end-to-end training and real-time inference speed of the object detectors. In this scenario, the object detector has a single neural network to extract the features for the regression of the bounding box and give the class probabilities without an auxiliary network for the region proposals. Recently, there are detectors that were developed to remove the requirement of defining anchor boxes during training. For instance, the Fully Convolutional One-Stage Object Detection (FCOS) is one of these models that, due to its nature, reduces all complicated computation related to anchor boxes, which can lead to an increase in inference time.

Learning using Privileged Information (LUPI). In human learning, the role of a teacher is crucial, guiding the students with additional information, such as explanations, comparisons, and so on [34]. In the LUPI setting, during the training, we have additional information provided by a teacher to help the learning procedure. Since the additional information is available at the training stage but not during the test time, we call it privileged information [34]. Re-

cently, [17] proposed the usage of privileged information to guide the variance of a Gaussian dropout. In a classification scenario, additional localization information is used, and its results show that it improves the generalization, requiring fewer samples for the learning process [17]. [21] designed a large-margin classifier using information bottleneck learning with privileged information for visual recognition tasks. In the object detection problem, [12] was the first work to present a modality hallucination framework, which incorporates the training RGB and Depth images, and during test time, RGB images are processed through the multi-modal framework to improve the performance of the detection. The modality hallucination network is responsible for mimicking depth mid-level features using RGB as input during the test phase. [20] used depth as privileged information for object detection with a Depth-Enhanced Deformable Convolution Network. In this work, we use the privileged information coming from a pre-trained RGB detector to improve the performance of the infrared detection. In practice, instead of destroying the information of the RGB detector by fine-tuning, we use the RGB detector as a guide for translating the IR input image into a new representation, which can help the RGB detector boost performance by enhancing the objects of interest.

Image Translation. The objective of image translation is to learn a mapping between two given domains such that images from the source domain can be translated to the target domain. In other words, the aim is to find a function $h_\vartheta : \mathcal{X}_\text{s} \rightarrow \mathcal{X}_\text{t}$ such that the distribution of images $h_\vartheta(\mathcal{X}_\text{s})$ in the translated domain is close to the distribution of images $\mathcal{X}_\text{t}$ in the target domain. Early methods rely on autoencoders (AEs) [11] and generative adversarial networks (GANs) [8] to learn cross-domain mapping. Unsupervised AE methods aim to learn a representation of the data by reconstructing the input data. GANs are a type of generative model that can learn to generate new data that is similar to the training data. More recently, diffusion models have gained popularity. They are capable of generating high-quality images but lack some properties for domain translation, like on CycleGANs. For improving models such as CycleGAN, techniques such as Contrastive Unpaired Translation (CUT) [23] and FastCUT [23] were developed. CUT is an image translation model based on maximizing mutual information of patches, which is faster than previous methods while providing results as good as others. On RGB/IR modalities, the InfraGAN [22] proposes an image-level adaptation using a model based on GANs, but for RGB to IR adaptation, with a focus on the quality of the generated images, thus optimizing image quality losses. Additionally, using image translation for object detection on RGB/IR using pre-train models, Herrmann et al. [10] used RGB object detectors without changing their parameters. The IR images are adapted to the RGB images using traditional computer vision pre-processing at the image level before applying it as input to the RGB object detector.

None of these methods provides an end-to-end way to directly train the image translation methods for detection applications. Furthermore, traditionally, they require more than one kind of data set composed of the original domain and the target domain. For instance, CycleGAN is based on adversarial loss, and U-net is based on reconstruction loss. Thus, if we have access to the already trained detector on the original domain, this knowledge can possibly be used during the learning of the translation network.

3. Proposed Method

Preliminary definitions. Let $\mathbf{x}_\text{i}$ be a given image with spatial resolution $W \times H$ and C channels. An object detector aims to output a set of N_reg object proposals, each represented as a bounding box $\mathbf{b}_\text{i,j} = (c, d, e, w, h)$, where (d, e) is the location of the top-left pixel of the bounding box for the j-th object, and w and h are the width and height of the object, respectively. Additionally, a classification label $c \in \{1, 2, ..., N_\text{cls}\}$ is assigned to each object of interest representing the region's class. In terms of optimization, such a task aims to maximize the detection accuracy, which typically is approximated through the average precision (AP) metric over all classes. Then, to train a detector, formally defined as the mapping $f_\theta \colon \mathbf{x}_\text{i} \rightarrow \hat{\mathbf{b}}_\text{i}$, a differentiable surrogate for AP metric is used, also known as the detection loss function, $\mathcal{L}_\text{det}(\mathbf{b}, \mathbf{x}; \theta)$.

The detection loss can be divided into two terms. The first one is the classification loss $\mathcal{L}_\text{cls}(\hat{y}_\text{c}, y_\text{c})$ responsible for learning the class label c. In this work, we use the cross-entropy loss function to assess the matching of bounding boxes categories $\mathcal{L}_\text{ce}(\hat{y}_\text{c}, y_\text{c}) = -\dfrac{1}{N_\text{cls}} \displaystyle\sum_{j=1}^{N_\text{cls}} y_{\text{c}_\text{j}} \log(p_\text{j})$, where N_cls is the total number of classes, and y_{c_j} is the class indicator function, i.e., $y_{\text{c}_\text{j}} = 1$ if c_j is the true class of the object, or $y_{\text{c}_\text{j}} = 0$ otherwise. The probability provided by the detector for each category j is p_j. To ensure the right positioning of the object, a second regression term $\mathcal{L}_\text{reg}(\hat{\mathbf{y}}_\text{b}, \mathbf{y}_\text{b})$ is used, being the $\mathcal{L}_\text{L1}(\hat{\mathbf{y}}_{\text{b}_\text{i}}, \mathbf{y}_{\text{b}_\text{i}}) = \sum_{i=1}^{N_\text{reg}} \left|\mathbf{y}_{\text{b}_\text{i}} - \hat{\mathbf{y}}_{\text{b}_\text{i}}\right|$ and $\mathcal{L}_\text{L2}(\hat{\mathbf{y}}_{\text{b}_\text{i}}, \mathbf{y}_{\text{b}_\text{i}}) = \sum_{i=1}^{N_\text{reg}} (\mathbf{y}_{\text{b}_\text{i}} - \hat{\mathbf{y}}_{\text{b}_\text{i}})^2$ losses the most commonly employed in the literature. Here N_reg is the number of bounding boxes on the image $\mathbf{x}_\text{i}$. Then, the final detection loss function can be defined in general terms as:

$$\begin{aligned}\mathcal{L}_\text{det}(\mathbf{x}, \mathbf{b}; \theta) = \mathcal{L}_\text{cls}(f_\theta(\mathbf{x}), c) \\ + \lambda \cdot \mathcal{L}_\text{reg}(f_\theta(\mathbf{x}), \mathbf{b}),\end{aligned} \tag{1}$$

where λ is a hyperparameter that controls the balance between the two terms, and θ is a vector containing the detec-

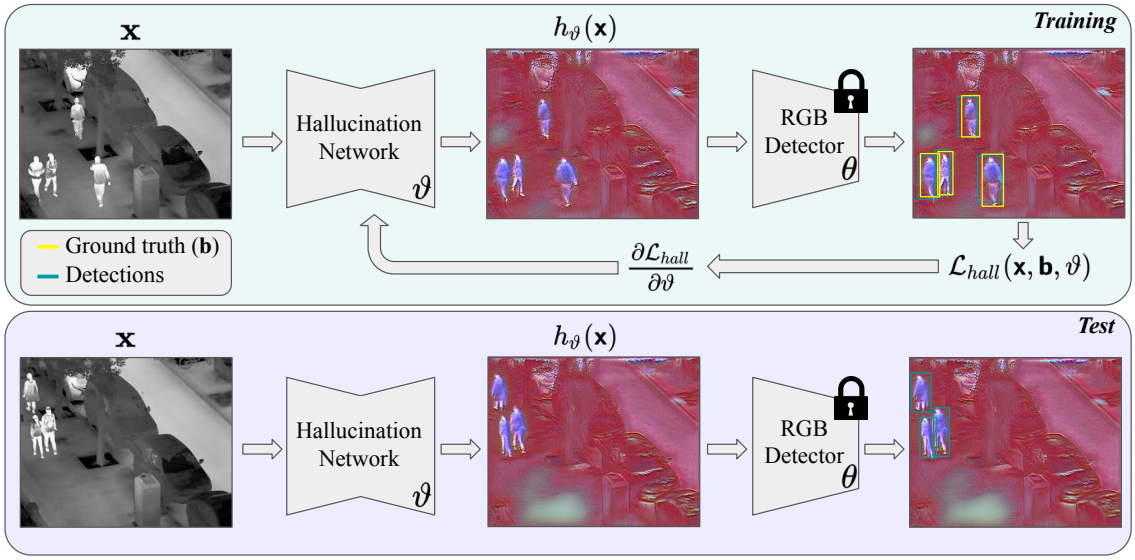


Figure 2. HalluciDet leverages privileged information for modality hallucination with pre-trained detectors. During training, the hallucination network learns how to use the privileged information encoded by the RGB detector to translate the IR image into a new hallucination modality representation. Then, during inference, the model provides better IR detection using the translated modality.

tor learnable parameters. The detectors used in this work use this general objective during their optimization process. However, they adapt each term to their specific architecture.

HalluciDet. Our goal is to generate a representation from an IR image that a given RGB detector can effectively process. Let $\mathcal{X} \subset \mathbb{R}^{W \times H}$ be the set of IR data containing N images. During the learning phase, a training dataset $S = \{(\mathbf{x}_\mathrm{i}, \mathbf{b}_\mathrm{i})\}$ is given such that $\mathbf{x}_\mathrm{i} \in \mathcal{X}$ is an IR image and $\mathbf{b}_\mathrm{i}$ is a set of bounding boxes as defined in the previous section. In addition, an RGB detector f_θ is also available. Then, a representation mapping is here defined as $h_\vartheta \colon \mathcal{X} \to \mathcal{R}$, where $\mathcal{R}$ is the representation space and ϑ are the learnable parameters of the translation model. Such a representation space, $\mathcal{R} \subset \mathbb{R}^{W \times H \times 3}$, is conditioned to the subset of plausible RGB images that are sufficient to obtain a proper response from the RGB detector f_θ. To find such a mapping we solve the optimization problem $\vartheta^* = \arg\min_\vartheta \mathcal{L}_{\text{hall}}(\mathbf{x}, \mathbf{b}; \vartheta)$ which implicitly uses the composition $(h_\vartheta \circ f_\theta)(\mathbf{x})$ to guide the intermediate representation.

Our proposed model, HalluciDet, comprises two modules: a hallucination network responsible for the image-to-image harmonization and a detector. The Hallucination network is based on U-net [29], but modified with attention blocks which are more robust for image translation tasks [6, 18]. For training the HalluciDet, we train the hallucination module and condition it with the detection loss, which is the only supervision necessary for guiding the hallucination training with respect to the privileged information of the pre-trained RGB detector. This phase is responsible for translating the hallucinated image to a new representation close to the RGB modality. Please note that this strategy helps the final model to perform well on the IR modality without changing the knowledge from the detector. Under this framework, the RGB detection performance remains the same since the detector's parameters θ are not updated during the adaptation learning. On the other hand, detections over IR images are obtained by adapting the input using the Hallucination network, followed by the evaluation over the RGB detector. As a side advantage, our model allows evaluating both modalities by providing the appropriate modality identifier during the forward pass, i.e., RGB or IR. Figure 2 depicts the training and evaluation process of an IR image using privileged information from the RGB detector.

The detector f_θ layers are frozen, thus preserving the prior knowledge, but the weights ϑ of the hallucination network h_ϑ are updated during the backward pass. The input minibatch is created with images from $\mathcal{X}$ set, leading to the hallucinated minibatch, which is then evaluated on f_θ to obtain the associated detections. To find the appropriate representation space, the hallucination loss $\mathcal{L}_{\text{hall}}(\mathbf{x}, \mathbf{b}, \vartheta)$ drives the optimization by updating only the hallucination network parameters. The representation space $\mathcal{R}$ is guided by $\mathcal{L}_{\text{hall}}$ to be closer enough to the RGB modality, which allows the detector to make successful predictions. As the representation is being learned with feedback from the frozen detector, it extracts the previous knowledge so that this new intermediate representation is tuned for the final detection task. The

proposed hallucination loss shares some similarities with the aforementioned detection loss but with the distinction of only updating the modality adaptation parameters:

$$\mathcal{L}_{\text{hall}}(\mathbf{x}, \mathbf{b}, \vartheta) = \mathcal{L}_{\text{cls}}(f_\theta(h_\vartheta(\mathbf{x})), c) + \lambda \cdot \mathcal{L}_{\text{reg}}(f_\theta(h_\vartheta(\mathbf{x})), \mathbf{b}) \quad (2)$$

Equation 2 is optimized w.r.t ϑ. We added the hyperparameter λ to weigh the contribution of each term and for numerical stability purposes.

4. Experimental results and analysis

Experimental Methodology. Hallucidet is evaluated on two different popular IR/RGB datasets, the LLVIP [15], and FLIR ADAS [7]. The LLVIP dataset is composed of $30,976$ images, in which $24,050$ ($12,025$ IR and $12,025$ RGB paired images) are used for training and $6,926$ for testing ($3,463$ IR and $3,463$ RGB paired images). For the FLIR, we used the sanitized and aligned paired sets provided by Zhang et al. [37], which have $10,284$ images, being $8,258$ for training ($4,129$ IRs and $4,129$ RGBs) and $2,026$ ($1,013$ IRs and $1,013$ RGBs) for test. We chose to utilize these paired IR/RGB datasets to ensure a fair comparison with other image-to-image translation techniques that employ reconstruction losses. In our experiments, we use 80% of the training set for training and the rest for validation. All results reported are on the test set. As for the FLIR dataset, we only used the person category. Initially, we have the RGB detector trained on the datasets using 5 different seeds. It's worth noting that this model starts with pre-trained weights from COCO [19]. Then with the RGB model trained, we use the model to perform the Hallucidet training. We tried ResNet_{50} as the backbone for the detectors and ResNet_{34} as the backbone for the Hallucination network. To ensure fairness we trained the detectors under the same conditions, i.e., data order, augmentations, etc. All the code is available at GitHub [1] for the reproducibility of the experiments. To develop the code, we used Torchvision models for the detectors and PyTorch Segmentation Models [14] for the U-Net architecture of the hallucination network. Additionally, we trained with PyTorch Lightning [5] training framework, evaluated the AP with TorchMetrics [4], and logged all experiments with WandB [1] logging tool.

Main Comparative Results. In Table 1, we investigate how our model behaved in comparison with standard image-to-image approaches and classical computer vision approaches that are normally used to reduce the distribution gap between IR and RGB. Furthermore, we highlight the impact of using the proposed $\mathcal{L}_{\text{hall}}$ loss to guide the representation. This is accomplished by comparing our approach with a U-Net that shares the same backbone as ours but employs a standard $\mathcal{L}_{\text{L1}}$ reconstruction loss. To guarantee comparability, we reproduce the experimental setting of [10] on our pipeline. We included basic pre-processing techniques that were shown to enhance IR performance on RGB models by Hermann et al. [10]. These techniques include a combination of blurring, histogram equalization, stretching, and inverting pixels. Furthermore, we included CycleGAN, which is a more powerful generative model compared with UNet. It is important to mention that training the CycleGAN is computationally more demanding than the Hallucidet. Additionally, due to the adversarial nature of the method, it does not ensure reliable convergence for the subsequent detection task. The CycleGAN was diverging with the same hyperparameters as [15] on the test set, so we tuned the hyperparameters and trained until the images became good qualitatively. Because CycleGAN introduces significant noise to the images as a result of its adversarial training, the detector's performance has notably decreased. This is particularly evident due to the increase in false positives. Given that our final goal is object detection, we selected FCOS, RetinaNet, and Faster R-CNN, each representing distinct categories within the universe of detection networks. We can see that straightforward approaches like inverting pixels for the IR and expanding it to three channels significantly enhance the initial performance of IR inputs on RGB detectors. As indicated in the table, our results demonstrate a significant improvement over previous image-to-image translation techniques in terms of detection performance. The most significant enhancement was observed in Faster R-CNN, where our proposal exhibited a remarkable 17% improvement compared to pixel inversion.

Hallucidet Visual Output. In Figure 3, we present a Hallucination image and compare it with both RGB and IR. The Hallucination emphasizes the person while smoothing the background, helping the detector to distinguish the regions of interest. In contrast to RGB, our method allows for easy person detection even in low-light conditions. However, IR images may introduce additional non-person-related information that could bias the detector. A visual comparison with FastCUT is also provided, revealing a correlation between the method's low performance and the high number of False Positives detected. It is important to note that while we show the Hallucination for representation demonstration, our main goal is on detection metrics. In the figure, the ground truth bounding box annotations are shown in yellow on the RGB images. The corresponding detections obtained from the IR data are presented in the following lines. It is important to note that we display the predicted detections on top of the intermediate representation for convenience.

[1] https://github.com/heitorrapela/HalluciDet.

Image-to-image translation	Learning strategy	AP@50↑		
		Test Set (Dataset: LLVIP)		
		FCOS	RetinaNet	Faster R-CNN
Blur [10]	-	42.59 ± 4.17	47.06 ± 1.99	63.05 ± 1.96
Histogram Equalization [10]	-	33.10 ± 4.64	36.45 ± 2.02	51.47 ± 4.03
Histogram Stretching [10]	-	38.55 ± 4.25	41.97 ± 1.39	57.69 ± 2.78
Invert [10]	-	53.62 ± 2.07	55.43 ± 2.03	71.83 ± 3.04
Invert + Equalization [10]	-	50.03 ± 2.44	52.57 ± 1.50	68.69 ± 2.73
Invert + Equalization + Blur [10]	-	50.58 ± 2.41	52.62 ± 1.36	68.91 ± 2.74
Invert + Stretching [10]	-	51.48 ± 2.17	52.87 ± 1.80	69.34 ± 3.07
Invert + Stretching + Blur [10]	-	51.54 ± 1.92	52.96 ± 1.80	69.59 ± 2.90
Parallel Combination [10]	-	50.18 ± 2.25	52.52 ± 1.39	68.14 ± 2.98
U-Net [29]	Reconstruction	42.94 ± 4.14	47.35 ± 1.92	63.23 ± 2.03
CycleGAN [39]	Adversarial	22.76 ± 1.94	27.04 ± 4.23	38.92 ± 5.09
CUT [24]	Contrastive learning	19.16 ± 2.10	21.61 ± 2.09	35.17 ± 0.32
FastCUT [24]	Contrastive learning	46.87 ± 2.28	52.39 ± 2.31	67.73 ± 2.14
HalluciDet (ours)	Detection	**63.28 ± 3.49**	**56.48 ± 3.39**	**88.34 ± 1.50**

Table 1. Performance comparison of models on IR images using LLVIP dataset [15]. The table showcases the impact of different approaches, including pixel manipulation techniques, U-Net, CycleGAN, CUT, FastCUT, and HalluciDet. The detectors were trained with RGB data and evaluated on IR. To make a fair comparison with our models, we decided to start with models that do not have strong data augmentation that could benefit one modality over the other.

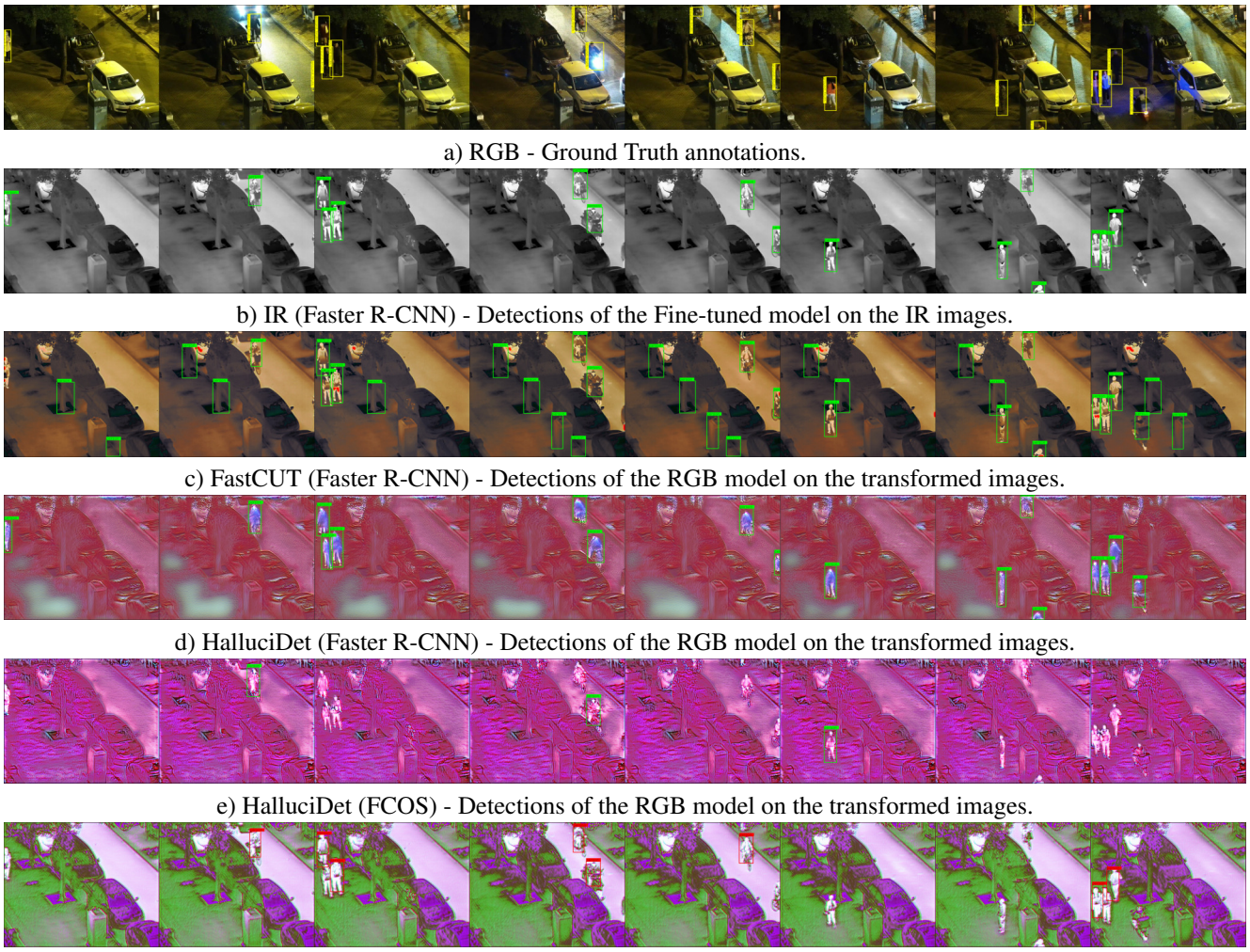
However, the actual inputs for HalluciDet approaches and FastCUT are IR images. A significant number of False Positives can be observed for FastCUT, while HalluciDet (FCOS) and HalluciDet (RetinaNet) exhibit a high number of False Negatives. The most accurate detection results are achieved with HalluciDet (Faster R-CNN), which demonstrates superior performance to the IR fine-tuned model in cases where the person's heat signature is not clearly evident, as seen in the last column. Additional figures can be found in the supplementary material.

Comparison with fine-tuning. For this experiment, we performed an evaluation of both RGB and fine-tuned IR detectors that were trained on the LLVIP and FLIR datasets. All methods from Table 2 were trained under the same experimental protocol using 3 different seeds.

Similar to the previous experiment, we utilized a detector from each family of methods, namely FCOS, RetinaNet, and Faster R-CNN. The provided results include the mean and standard deviation of the AP on the test set. In this experiment, we compare three different approaches to adapt a model trained on RGB images to IR. As baseline we consider the case of No Adaptation, in which the model is used directly on IR images. Then, we consider the case in which a model is adapted to the IR data with normal fine-tuning, which is the most common way of adaptation when annotations are available. Finally, we train our HalluciDet to generate a new representation of the image for the RGB detector.

As seen in Table 2, in all cases, the fine-tuned IR model outperformed the RGB detector over the IR modality, as expected. In the tables, we also observe a significant improvement in the performance of HalluciDet compared to the performance achieved through fine-tuning for Faster R-CNN. This improvement aligns with the quality of the representation observed in Figure 3, where confusing factors, such as car heat, have been removed from the image. A marginal improvement was observed with center point-based architectures like FCOS for the LLVIP dataset, although a higher difference in AP could be observed for the FLIR dataset. On the other hand, the results using RetinaNet didn't exhibit much consistency; the AP was significantly worse than that achieved through fine-tuning for the LLVIP dataset. Once again, this is consistent with the observed representation lacking the necessary discriminative information to detect people in the image.

Hallucidet with different backbones. In Table 3, we investigated various encoder backbones for the Hallucination network. The presented results include two MobileNet and two ResNets with different widths. Additional outcomes for alternative backbones are included in the supplementary material. In all cases, the model consistently improves upon the performance of the fine-tuned IR model. Notably, even in models with a reduced number of parameters, such as MobileNet$_{v2}$ with less than 7 million additional parameters,



a) RGB - Ground Truth annotations.

b) IR (Faster R-CNN) - Detections of the Fine-tuned model on the IR images.

c) FastCUT (Faster R-CNN) - Detections of the RGB model on the transformed images.

d) HalluciDet (Faster R-CNN) - Detections of the RGB model on the transformed images.

e) HalluciDet (FCOS) - Detections of the RGB model on the transformed images.

f) HalluciDet (RetinaNet) - Detections of the RGB model on the transformed images.

Figure 3. Illustration of a sequence of 8 images of LLVIP dataset. The first row is the RGB modality, then the IR modality, followed by FastCUT and different representations created by HalluciDet over various detectors.

the gain remains consistent at nearly 5%.

Hallucidet with a different number of training samples. For the LLVIP dataset, in Figure 4, we explored various quantities of training samples for our method, ranging from 1% to 100%. Notably, only 30% of the data was sufficient for HalluciDet to achieve comparable performance to the fine-tuned Faster R-CNN with the complete dataset. For the FLIR dataset, in Figure 5, the trend to reduce the number of training samples and improve over the fine-tuning is still true, but in this case, around 70% of the training samples. The different characteristics related to the exact number of training samples with respect to the dataset are due to the number of different environment changes on the datasets. For the LLVIP, we do not have a big shift in the images because the cameras are fixed in a surveillance context. In the case of FLIR, the variance of the images is higher due to the different capture settings; with the focus on autonomous driving, the camera moves inside a car, which changes the background consistency and introduces more variance to the dataset.

5. Conclusion

In this work, we provided a framework that uses privileged information of an RGB detector to perform the image-to-image translation from IR. The approach involves utilizing a Hallucination network to generate intermediate representations from IR data, which are then directly input into an RGB detector. An appropriate loss function was also proposed to lead the representation into a space that allows

Method	AP@50↑		
	Test Set IR (Dataset: LLVIP)		
	No Adaptation	Fine-tuning	HalluciDet
FCOS	47.12 ± 4.32	63.79 ± 0.48	**64.85 ± 1.46**
RetinaNet	50.63 ± 3.22	**76.26 ± 0.75**	56.78 ± 3.85
Faster R-CNN	71.51 ± 1.16	84.94 ± 0.15	**90.92 ± 0.20**
	Test Set IR (Dataset: FLIR)		
	No Adaptation	Fine-tuning	HalluciDet
FCOS	38.52 ± 0.79	42.22 ± 1.04	**49.18 ± 0.99**
RetinaNet	44.13 ± 2.01	47.87 ± 2.21	**49.01 ± 4.08**
Faster R-CNN	55.85 ± 1.19	61.48 ± 1.55	**70.90 ± 1.35**

Table 2. AP performance for various models following distinct training approaches on two datasets of LLVIP [15] (top half) and FLIR [9] (bottom half): starting from COCO pre-training and fine-tuning on the RGB data shown as (No Adaptation) and fine-tuning on the IR data shown as (Fine-tuning). In the case of HalluciDet, the trained RGB detector serves as the initial point, with the subsequent optimization of the Hallucination network using the IR data. The reported performance is exclusive to the person category.

Method		Params.	AP@50↑
Faster R-CNN		41.3 M	84.83
HalluciDet	MobileNet_{v3s}	+ 3.1 M	85.20
	MobileNet_{v2}	+ 6.6 M	89.73
	ResNet_{18}	+ 14.3 M	90.42
	ResNet_{34}	+ 24.4 M	90.65

Table 3. Comparison of the number of parameters for different Hallucination Network backbones vs. AP@50 on the LLVIP dataset with the Faster R-CNN detector.

for the enhancement of the target category's importance.

In our experiments, we demonstrate that hallucination networks can be helpful for modality adaptation by obtaining an intermediate representation that effectively supports accurate responses in the object detection task. The proposed approach showed particular effectiveness for the two-stage detector Faster R-CNN, resulting in a reduction of non-person-related information. This reduction in background clutter had a positive effect on minimizing the number of False Positives, surpassing the performance of standard fine-tuning on IR data. The comparison with methods from the literature for image-to-image translation highlighted the significance of guiding the representation to achieve successful detections. Our Hallucidet demonstrated a significant performance improvement compared to the other methods. Finally, the proposed framework offers the additional advantage of maintaining performance in the RGB task, which is beneficial for applications requiring accurate responses in both modalities.

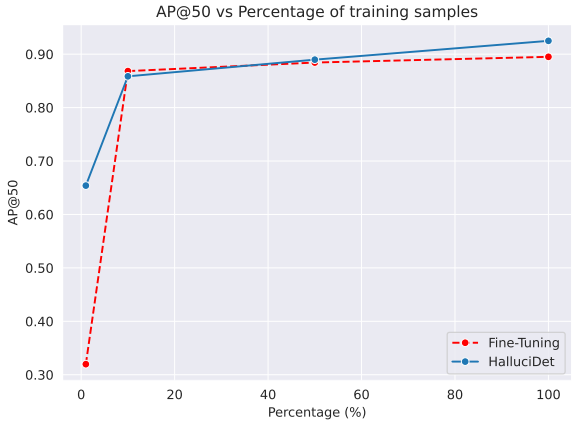


Figure 4. AP@50 vs. training samples percentages. The figure shows the AP@50 over the LLVIP test set using various amounts of training samples for the HalluciDet Faster R-CNN.

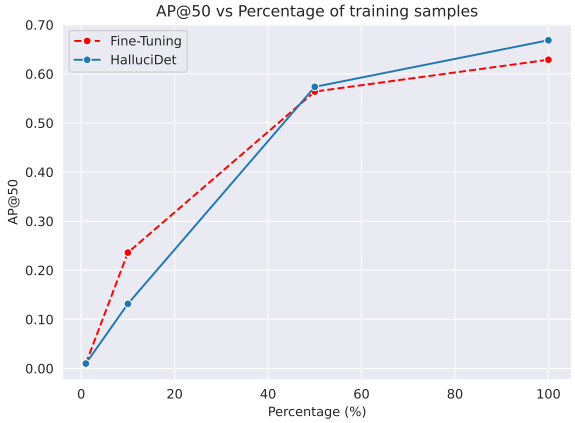


Figure 5. AP@50 vs. training samples percentages. The figure shows the AP@50 over the FLIR test set using various amounts of training samples for the HalluciDet Faster R-CNN. Notably, 70% of the data was sufficient for HalluciDet to achieve comparable performance to the fine-tuned Faster R-CNN with the complete dataset.

Acknowledgments: This work was supported by Distech Controls Inc., the Natural Sciences and Engineering Research Council of Canada, the Digital Research Alliance of Canada, and MITACS.

References

[1] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com.

[2] Jianguo Chen, Kenli Li, Qingying Deng, Keqin Li, and S Yu Philip. Distributed deep learning model for intelligent video surveillance systems with edge computing. *IEEE Transactions on Industrial Informatics*, 2019.

[3] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3819–3824. IEEE, 2018.

[4] Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh Jha, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. Torchmetrics-measuring reproducibility in pytorch. *Journal of Open Source Software*, 7(70):4101, 2022.

[5] William Falcon and The PyTorch Lightning team. PyTorch Lightning, Mar. 2019.

[6] Tongle Fan, Guanglei Wang, Yan Li, and Hongrui Wang. Ma-net: A multi-scale attention network for liver and tumor segmentation. *IEEE Access*, 8:179656–179665, 2020.

[7] Teledyne Flir. Free flir thermal dataset for algorithm training. teledyne flir llc all rights reserved, 2021.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[9] FA Group et al. Flir thermal dataset for algorithm training, 2018.

[10] Christian Herrmann, Miriam Ruf, and Jürgen Beyerer. Cnn-based thermal infrared person detection by domain adaptation. In *Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*, volume 10643, page 1064308. International Society for Optics and Photonics, 2018.

[11] Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and helmholtz free energy. *Advances in neural information processing systems*, 6, 1993.

[12] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 826–834, 2016.

[13] Yu Huang and Yue Chen. Survey of state-of-art autonomous driving technologies with deep learning. In *2020 IEEE 20th international conference on software quality, reliability and security companion (QRS-C)*, pages 221–228. IEEE, 2020.

[14] Pavel Iakubovskii. Segmentation models pytorch, 2019.

[15] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3496–3504, 2021.

[16] Xiangyin Kong and Zhiqiang Ge. Deep learning of latent variable models for industrial process monitoring. *IEEE Transactions on Industrial Informatics*, 18(10):6778–6788, 2021.

[17] John Lambert, Ozan Sener, and Silvio Savarese. Deep learning under privileged information using heteroscedastic dropout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8886–8895, 2018.

[18] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018.

[19] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[20] Yan Liu, Zhijie Zhang, and Liqing Zhang. Depth privileged object detection with depth-enhanced dcn. In *International Conference on Neural Information Processing*, pages 438–446. Springer, 2021.

[21] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Information bottleneck learning using privileged information for visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1496–1505, 2016.

[22] Mehmet Akif Özkanoğlu and Sedat Ozer. Infragan: A gan architecture to transfer visible images to infrared domain. *Pattern Recognition Letters*, 155:69–76, 2022.

[23] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020.

[24] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, 2020.

[25] Harry A Pierson and Michael S Gashler. Deep learning in robotics: a review of recent research. *Advanced Robotics*, 31(16):821–835, 2017.

[26] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[30] Stamatios Samaras, Eleni Diamantidou, Dimitrios Ataloglou, Nikos Sakellariou, Anastasios Vafeiadis, Vasilis Magoulianitis, Antonios Lalas, Anastasios Dimou, Dimitrios Zarpalas, Konstantinos Votis, et al. Deep learning on multi sensor data for counter uav applications—a systematic review. *Sensors*, 19(22):4837, 2019.

[31] Jack Stilgoe. Machine learning, social learning and the governance of self-driving cars. *Social studies of science*, 48(1):25–56, 2018.

[32] Karasawa Takumi, Kohei Watanabe, Qishen Ha, Antonio Tejero-De-Pablos, Yoshitaka Ushiku, and Tatsuya Harada. Multispectral object detection for autonomous vehicles. In

Proceedings of the on Thematic Workshops of ACM Multimedia 2017, pages 35–43, 2017.

[33] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[34] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

[35] Siwei Yang, Shaozuo Yu, Bingchen Zhao, and Yin Wang. Reducing the feature divergence of rgb and near-infrared images using switchable normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 46–47, 2020.

[36] Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. Dive into deep learning. *arXiv preprint arXiv:2106.11342*, 2021.

[37] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, 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.

[38] Huaizhong Zhang, Chunbo Luo, Qi Wang, Matthew Kitchin, Andrew Parmley, Jesus Monge-Alvarez, and Pablo Casaseca-De-La-Higuera. A novel infrared video surveillance system using deep learning based techniques. *Multimedia tools and applications*, 77:26657–26676, 2018.

[39] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.