Hallucinating RGB Modality for Person Detection Through Privileged Information

Heitor Rapela Medeiros, Fidel A. Guerrero Peña, Masih Aminbeidokhti
Eric Granger, Marco Pedersoli
LIVIA, Dept. of Systems Engineering, ETS Montreal, Canada
heitor.rapela-medeiros.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. 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. 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 pretrained RGB detector.

1. Introduction

Despite the impressive performance of deep learning (DL) models, their effectiveness can significantly deteriorate when applied to modalities that were not present during the training [1, 12]. For example, a model trained on RGB images may not perform well on IR images during testing [14]. 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. In Figure 1, the detections of our method are illustrated and compared with competitors. 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:**

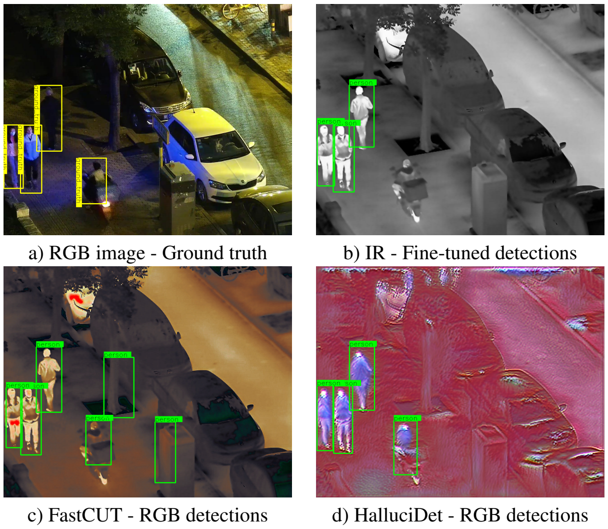

a) RGB image - Ground truth
b) IR - Fine-tuned detections
c) FastCUT - RGB detections
d) HalluciDet - RGB detections

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

(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), we compare HalluciDet against various image-to-image translation. Our approach is seen to improve detection accuracy on the IR modality by incorporating privileged information from RGB.

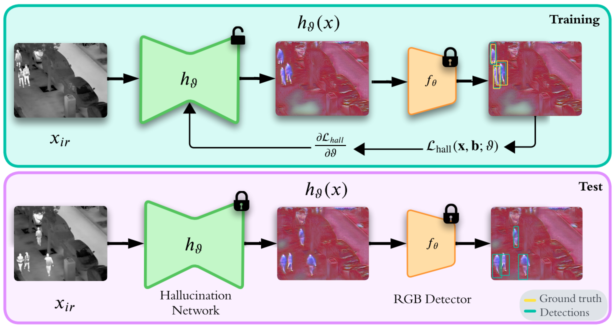


Figure 2. HalluciDet leverages privileged information for modality hallucination with pre-trained detectors. During training, the hallucination network learns how to use the privileged information encoded by the RGB detector to translate the IR image into a new hallucination modality representation.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

2. Related Work

Object detection: 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. The one-stage detectors mainly focus on end-to-end training and real-time inference speed of the object detectors. RetinaNet [6] is a one-stage detector with focal loss for better detection of hard objects. FCOS [11] is a one-stage detector, which reduces all complicated computation related to anchor boxes, which can lead to an increase in inference time. The Faster R-CNN [9] is a two-stage detector with a Region Proposal Network and then a classification part. **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 [13]. 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 [13]. In this work, we use the privileged information coming from a pre-trained RGB detector to improve the performance of the infrared detection. **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. A common approach is CycleGAN, which a GAN for translation between two domains, then techniques such as Contrastive Unpaired Translation (CUT) [7] and FastCUT [7] 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.

3. Proposed Method

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.[1] During the learning phase, a training dataset $S = \{(\mathbf{x}_\text{i}, \mathbf{b}_\text{i})\}$ is given such that $\mathbf{x}_\text{i} \in \mathcal{X}$ is an IR image and $\mathbf{b}_\text{i}$ is a set of bounding boxes. In addition, an RGB detector f_θ is also available. Then, a representation mapping is here defined as $h_\vartheta : \mathcal{X} \to \mathcal{R}$, 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 [10], but modified with attention blocks which are more robust for image translation tasks [2, 5]. 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

[1] The term Hallucination was used for the model's power to provide a pixel color space for the intermediated image that was useful for the detector even though there is no constraint on the colors of the pixels.

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:

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

Equation 1 is optimized w.r.t ϑ. We added the hyperparameter λ to weigh the contribution of each term and for numerical stability purposes. Where $\mathcal{L}_{\text{cls}}$ is the cross-entropy between the classes detected on the intermediated image and the GT, and $\mathcal{L}_{\text{reg}}$ is l1 loss between the detected bound boxes and the GT boxes.

Image-to-image translation	Learning strategy	AP@50↑		
		Test Set (Dataset: LLVIP)		
		FCOS	RetinaNet	Faster R-CNN
U-Net [10]	Reconstruction	42.94 ± 4.14	47.35 ± 1.92	63.23 ± 2.03
CycleGAN [15]	Adversarial	22.76 ± 1.94	27.04 ± 4.23	38.92 ± 5.09
CUT [8]	Contrastive learning	19.16 ± 2.10	21.61 ± 2.09	35.17 ± 0.32
FastCUT [8]	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 [4]. 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.

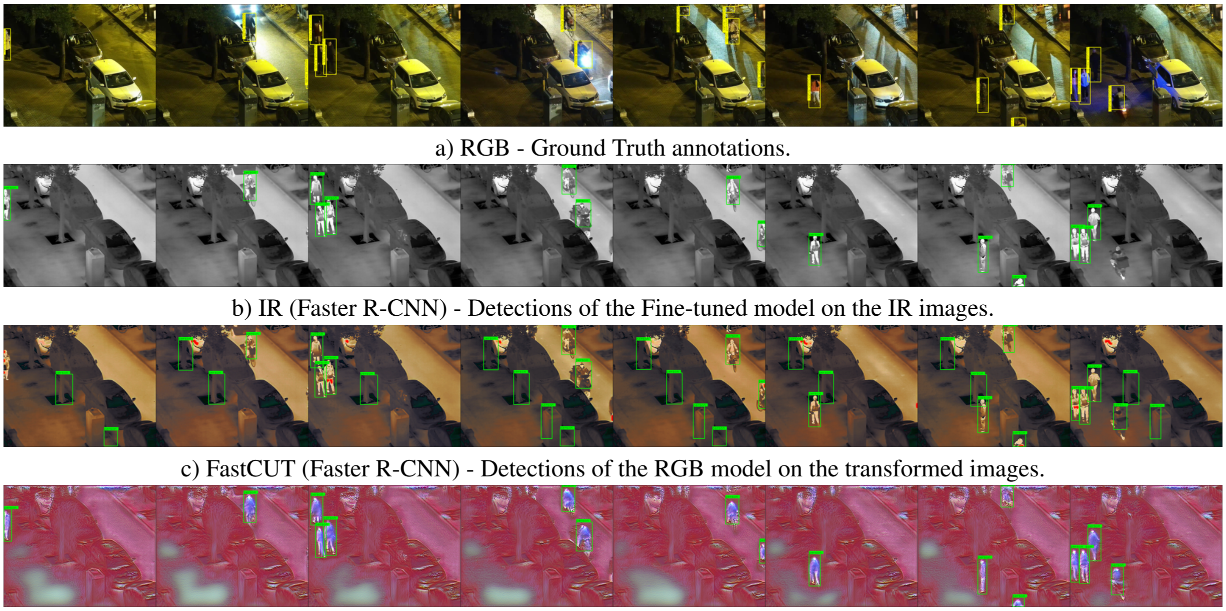
4. Experimental results and analysis

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. 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. 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. As indicated in the table, our results demonstrate a significant improvement over previous image-to-image translation techniques in terms of detection performance. **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 3, 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. 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.

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 [4] (top half) and FLIR [3] (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).

Comparison with fine-tuning: We performed an evaluation of both RGB and fine-tuned IR detectors that were



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.

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.

trained on the LLVIP and FLIR datasets, see Table 2. 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. Then, we consider IR fine-tuning, which is the most common way of adaptation when annotations are available. Finally, 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 table, 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.

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.

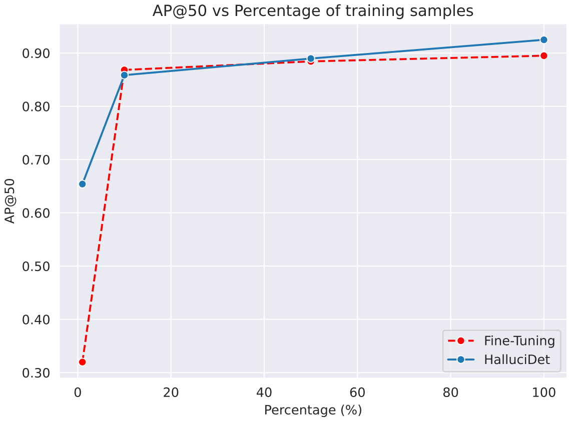


Figure 4. AP@50 vs. training samples percentages. The figure shows the AP@50 over the LLVIP test set using various amounts of training samples for the HalluciDet Faster R-CNN.

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 for the enhancement of the target category's importance. 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.

References

[1] 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. 1

[2] 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. 2

[3] FA Group et al. Flir thermal dataset for algorithm training, 2018. 3

[4] 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. 3

[5] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018. 2

[6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[7] 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. 2

[8] 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. 3

[9] 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. 2

[10] 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. 2, 3

[11] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: A simple and strong anchor-free object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4): 1922–1933, 2020. 2

[12] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. 1

[13] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009. 2

[14] 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. 1

[15] 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. 3