Supplementary Material: 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

In this supplementary material, we provide additional information to reproduce our work. The source code[1] is publicly provided. Here, we provide ablation on the hyperparameter of the HalluciDet loss, qualitative examples of the obtained detections, and additional results.

A. Ablation of hyperparameters λ for HalluciDet

In this section, we show the sensitivity of HalluciDet to the hyperparameters during training. For these experiments, as we did not want to have the influence of data augmentation on the pipeline, we removed the data augmentations that could benefit the starting detector and also the HalluciDet. Thus, the results in the main manuscript are the results with the detector using transformations such as color jitter and horizontal flip, and the same transformations were used for training the HalluciDet and respective baselines. In this ablation, we focused on the balancing of λ, so for this case, we kept both detectors and HalluciDet without data augmentations.

The cost function of the hallucination network $\mathcal{L}_{hall}$ (Equation 1) contains three terms: regression loss, classification loss, and other losses. Here, the other loss terms are dependent on the detection method used, e.g., for Faster-RCNN $\mathcal{L}_* = \mathcal{L}_{rpn} + \mathcal{L}_{obj}$, where the regression loss $\mathcal{L}_{rpn}$ is applied to the region proposal network, and $\mathcal{L}_{obj}$ is the object/background classification loss.

$$\mathcal{L}_{\text{hall}} = \lambda_{cls} \cdot \mathcal{L}_{\text{cls}} + \lambda_{reg} \cdot \mathcal{L}_{\text{reg}} + \lambda_* \cdot \mathcal{L}_* \qquad (1)$$

As shown on Table 1 of this supplementary materials, the HalluciDet ablation study was divided into different ways of balancing the regression and classification parts of the loss. In practice, it is better to use both components (regression and classification), but we recommend prioritizing the regression part for optimal balance.

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

B. Hallucidet and additional results on FLIR

Similar to the main manuscript, we added additional ablations with respect to the FLIR dataset.

Hallucidet with a different encoder. Similar to the main manuscript, we provided a study on the different backbones of the hallucination network encoder but focused on the FLIR dataset. The results show a similar trend, in which models with more capacity in terms of parameters can learn more robust representations for the test set distribution, thus increasing the AP@50.

C. Qualitative analysis of Hallucidet Detections

In this section, we provided an additional sequence of batch images, similar to the main manuscript. Here, we can find more than one batch of 8 images for the LLVIP dataset (Figure 1), and then two batches of 8 images each for the FLIR dataset (Figure 2, Figure 3). Thus, the trend and explanations for detections remain the same as those described in the main manuscript.

Processing time comparison: In terms of trade between more parameters that can increase the speed for processing and the performance of the detection, we highlight some important discussion about it. For the models classified as nonlearning methods, there is no increase in the inference speed and the training part, but they have lower detection performance. For the models that are learning in the input

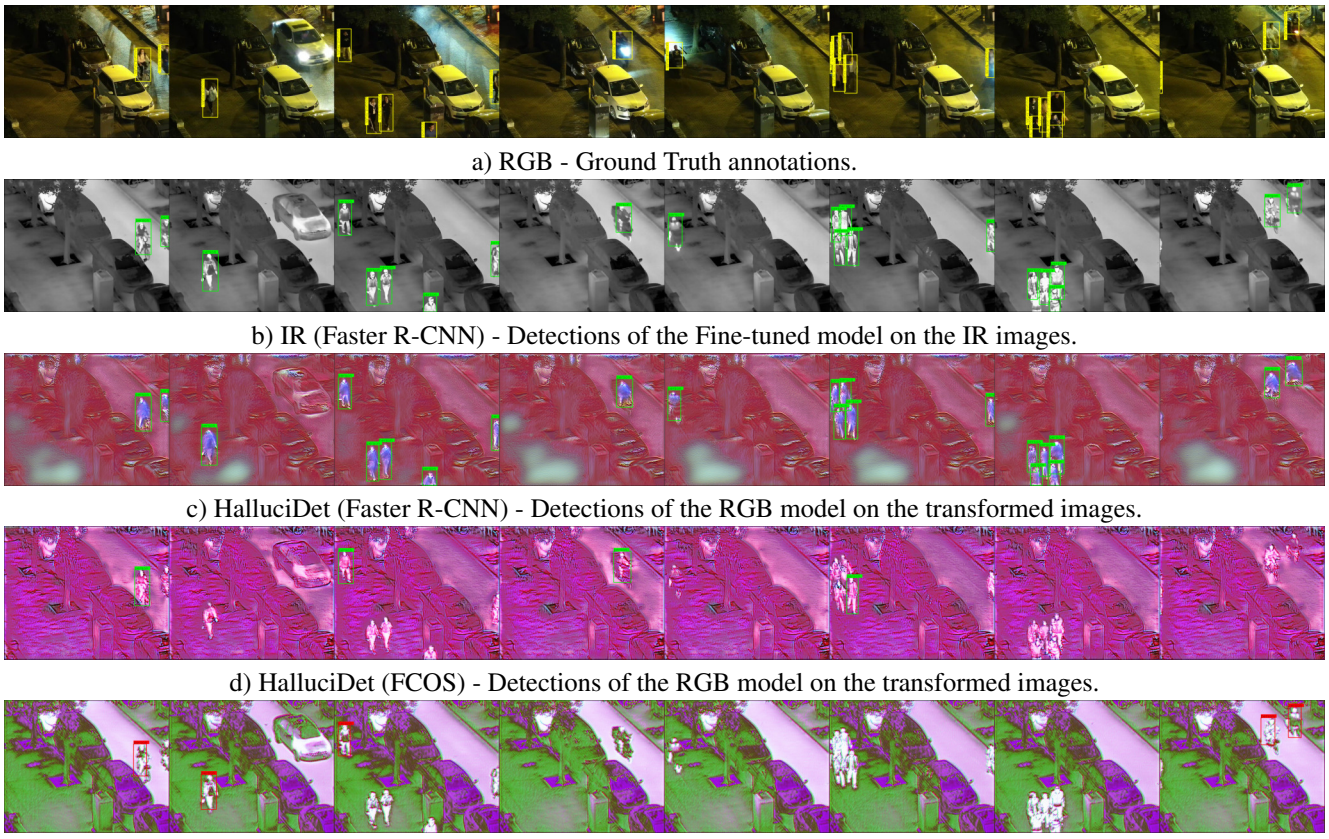
Method	Ablation (Loss Weight)	AP@0.5↑
		Test Set (Dataset: LLVIP) $\mathcal{X}^{IR}$
HalluciDet (RetinaNet)	$\lambda_{cls} = 0.0$, $\lambda_{reg} = 1.0$	65.81
	$\lambda_{cls} = 1.0$, $\lambda_{reg} = 0.0$	60.77
	$\lambda_{cls} = 0.01$, $\lambda_{reg} = 0.1$	68.58
	$\lambda_{cls} = 0.1$, $\lambda_{reg} = 0.01$	60.03
HalluciDet (FCOS)	$\lambda_{cls} = 0.0$, $\lambda_{reg} = 1.0$, $\lambda_{box_{cnt}} = 1.0$	63.01
	$\lambda_{cls} = 1.0$, $\lambda_{reg} = 0.0$, $\lambda_{box_{cnt}} = 0.0$	60.92
	$\lambda_{cls} = 0.01$, $\lambda_{reg} = 0.1$, $\lambda_{box_{cnt}} = 0.1$	65.02
	$\lambda_{cls} = 0.1$, $\lambda_{reg} = 0.01$, $\lambda_{box_{cnt}} = 0.01$	64.59
HalluciDet (Faster R-CNN)	$\lambda_{cls} = 0.1$, $\lambda_{obj} = 0.1$, $\lambda_{reg} = 0.01$, $\lambda_{RPNbox_{reg}} = 0.01$	85.35
	$\lambda_{cls} = 0.01$, $\lambda_{obj} = 0.01$, $\lambda_{reg} = 0.1$, $\lambda_{RPNbox_{reg}} = 0.1$	88.72
	$\lambda_{cls} = 1.0$, $\lambda_{obj} = 1.0$, $\lambda_{reg} = 0.0$, $\lambda_{RPNbox_{reg}} = 0.0$	83.97
	$\lambda_{cls} = 0.0$, $\lambda_{obj} = 0.0$, $\lambda_{reg} = 1.0$, $\lambda_{RPNbox_{reg}} = 1.0$	84.08

Table 1. Comparison between different weights on the losses terms. In this table, the models are started frozen from RGB, the same as reported in the paper. Then, the hallucination network is trained with different lambda values to see its impacts on the model's performance. Results over LLVIP test set.

Method		Params.	AP@50↑
Faster R-CNN		41.3 M	61.48
HalluciDet	MobileNet_{v3s}	+ 3.1 M	53.62
	MobileNet_{v2}	+ 6.6 M	67.74
	ResNet_{18}	+ 14.3 M	68.56
	ResNet_{34}	+ 24.4 M	71.58

Table 2. Comparison of the number of parameters for different Hallucination Network backbones vs. AP@50 on the FLIR dataset with the Faster R-CNN detector.

space, such as image translation methods like CycleGAN or FastCUT, given the same backbone network, HalluciDet has faster training and equal inference time to the deep learning baselines, and we can improve the detection performance.



a) RGB - Ground Truth annotations.

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

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

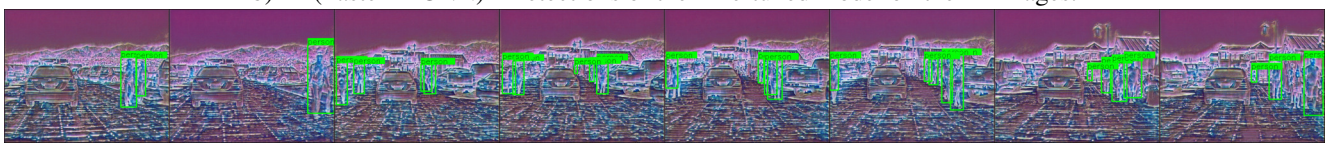
d) HalluciDet (FCOS) - Detections of the RGB model on the transformed images.

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

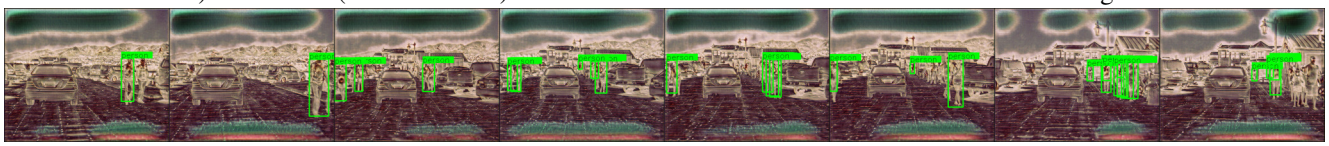
Figure 1. Illustration of a sequence of 8 images of LLVIP dataset. The first row is the RGB modality, then the IR modality, followed by different representations created by HalluciDet over various detectors.

a) RGB - Ground Truth annotations.

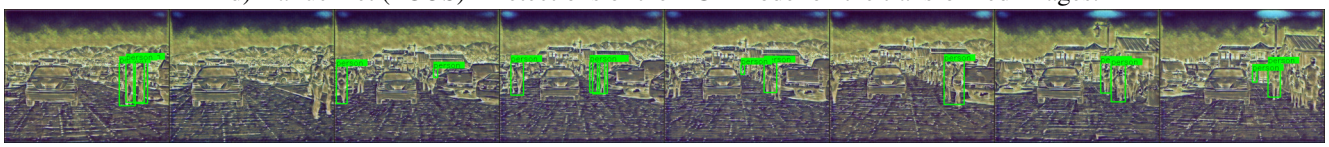
b) IR (Faster R-CNN) - Detections of the Fine-tuned model on the IR images.



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



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

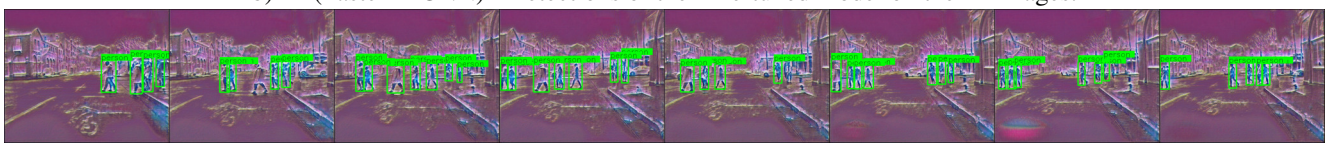


e) HalluciDet (RetinaNet) - Detections of the RGB model on the transformed images.

Figure 2. Illustration of a sequence of 8 images of FLIR dataset. The first row is the RGB modality, then the IR modality, followed by different representations created by HalluciDet over various detectors.

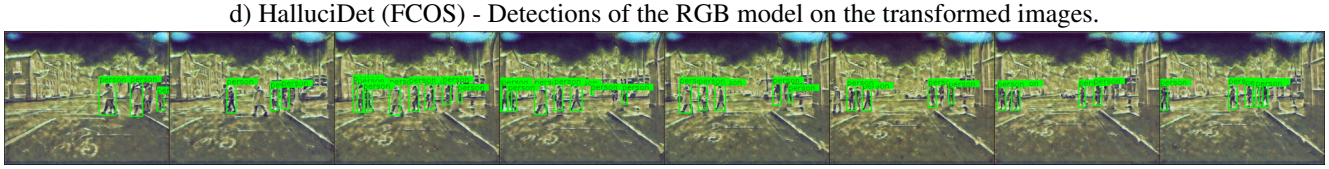
a) RGB - Ground Truth annotations.

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



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

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



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

Figure 3. Another sequence of 8 images of FLIR dataset. The first row is the RGB modality, then the IR modality, followed by different representations created by HalluciDet over various detectors.