Pixel Invisibility: Detect Object Unseen in Color Domain

Yongxin Wang[1][a] and Duminda Wijesekera[2][b]
[1]*Department of Computer Science, George Mason University, 4400 University Dr, Fairfax, VA 22030, U.S.A.*
[2]*Department of Computer Science and Cyber Security Engineering, George Mason University, 4400 University Dr, Fairfax, VA 22030, U.S.A.*

Keywords: Color-thermal Image Pairs, Pixel Invisibility, Cross-modality Distillation.

Abstract: Deep neural networks have been very successful in image recognition. In order for those results to be useful for driving automatons require quantifiable safety guarantees during night, dusk, dawn, glare, fog, rain and snow. In order to address this problem, we developed an algorithm that predicts a pixel-level invisibility map for color images that does not require manual labeling - that computes the probability that a pixel/region contains objects that are invisible in color domain, during light challenged conditions such as day, night and fog. We do so by using a novel use of cross modality knowledge distillation from color to thermal domain using weakly-aligned image pairs obtained during daylight and construct indicators for the pixel-level invisibility by mapping both the color and thermal images into a shared space. Quantitative experiments show good performance of our pixel-level invisibility masks and also the effectiveness of distilled mid-level features on object detection in thermal imagery.

1 INTRODUCTION

Object detection in imagery has rapidly improved since publication of large scale data sets (Deng et al., 2009; Everingham et al., 2010; Geiger et al., 2013; Lin et al., 2014; Yu et al., 2018) and powerful baseline systems like two-stage detectors, Fast/Faster/Mask R-CNN (Girshick, 2015; Ren et al., 2015; He et al., 2017) and one-stage detectors, such as YOLO (Redmon et al., 2016; Redmon and Farhadi, 2017; Redmon and Farhadi, 2018), SSD (Liu et al., 2016), RetinaNet (Lin et al., 2017). One important issue in using object detectors for driving automatons are guarantees on high confidence to ensure that navigable regions are free of obstructing objects during all operational weather and lighting conditions. Failing to detect object-free regions or provide warning signals, for example when crossing pedestrians or vehicles can lead to safety violations. While the performance of object detectors is improving, they cannot be guaranteed never to make mistakes (Rahman et al., 2019). Thus reliable vision systems should account for *knowing when they cannot recognize objects* in addition to providing high detection accuracy. Our work attempts to address this problem by predicting a so-called *pixel-level invisibility map for color images* without manual labeling. Invisibility maps can be used by a system to trust detected results of some regions over others in an image or in signal warning messages sent to a driving automation. When multiple sensors are used to increase detection reliability, invisibility maps will guide the sensor fusion process and improve the overall performance for detection.

We define an *invisibility mask* for one image as the likelihood that a region or pixel contains objects invisible in that domain. That is, the likelihood of one pixel or region contributing to false negatives in object detectors due to poor visibility. Regions of color images during good daylight obtain low invisibility scores because visible light of enough energy is reflected to the camera by objects on the spot. Though dark regions of images in the night or obscure regions of images in the fog will have high invisibility scores. One straightforward approach to create invisibility maps is to create a large labeled data set where every pixel in the image is labeled with an invisibility score - very expensive to collect. If one were to do so by hand, human judgements for predictions about the invisible area for an image is highly subjective, but the detection done in software. Instead, our method predicts the invisibility score for every pixel in the image without laborious human labeling by proposing

[a] https://orcid.org/0000-0001-8343-4557
[b] https://orcid.org/0000-0002-7122-3055

Wang, Y. and Wijesekera, D.
Pixel Invisibility: Detect Object Unseen in Color Domain.
DOI: 10.5220/0010479002010210
In *Proceedings of the 7th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2021)*, pages 201-210
ISBN: 978-989-758-513-5

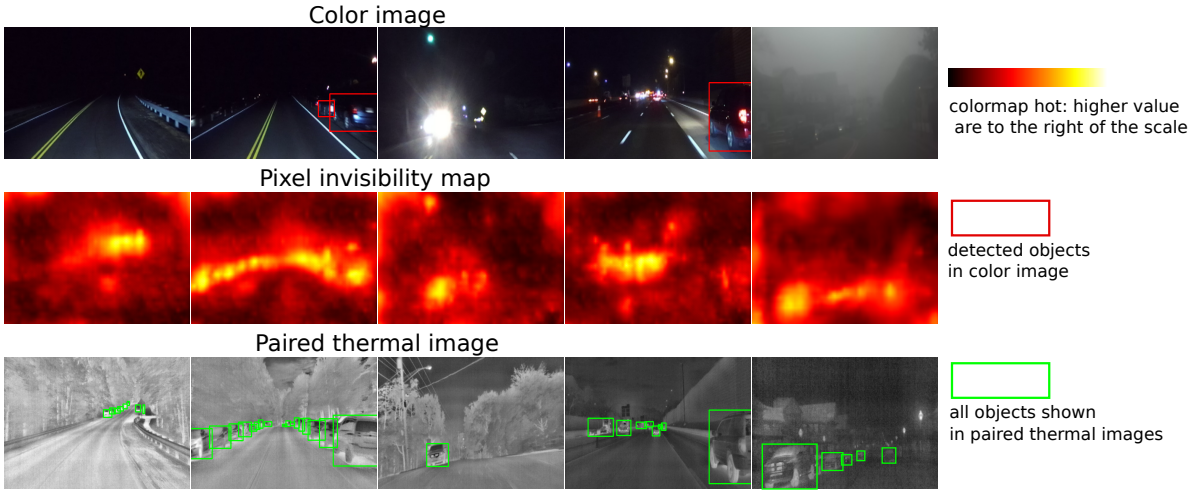

Figure 1: Our method predicts an invisibility map for each pixel (2^{nd} row) given only the color image (1st row). Left to right: It can handle distant objects, highly occluded and cluttered objects, objects with strong glare, multiple objects from complex scenes like working zones on highways and objects in the fog. Color detectors can only detect objects (bounding boxes in 1st row) that are visible to its spectrum, while there are many more objects in the scene (3rd row). This could cause disastrous consequences where safety is crucial if such missed detection is trusted. Our pixel-level invisibility map indicates how much the detection results from color images may be trusted, both for false negatives and false positives.

a novel use of cross modality knowledge distillation and the generation of well-aligned image pairs between color and thermal images. There is a difference between creating invisibility maps and uncertainty estimation for detectors; the former attempts to estimate the sensor limitation for each pixel in the image, the latter attempts to estimate the failure of the proposed detector itself.

Cross modality knowledge distillation (Gupta et al., 2016) is also called supervision transfer as a means of avoiding labeling large scale data sets for certain modalities. Given paired images from two modalities, intermediate-level representations are transferred from richly annotated color domain to other modalities with limited labeled data sets such as depth images (Gupta et al., 2016), sound (Aytar et al., 2016) and thermal images. The novelty of our invisibility mask is in mapping images from two different modalities into one shared space by using the supervision transfer from color to thermal images of the daytime and then approximating the invisibility of color images through perceptual distances between two modalities in challenging lighting conditions including dusk, dawn, dark nights and fog.

Knowledge distillation (Gupta et al., 2016) requires that the two modalities are presented in a paired fashion, especially in a well-aligned manner for object detection and for our pixel-level invisibility masks. Here, well-aligned image pairs are the ones where the corresponding pixels in the paired images are located at the same positions in the respective image planes. The raw image pairs captured by color and thermal sensors have large displacements (Zhang et al., 2019) (Li et al., 2018), which come from (1) Internal camera attribute differences such as focal length, resolution and lens distortion, (2) External transformations like pose differences and (3) Time differences from exposure time and shutter speed. We address the first two disparities by image registration using a known pattern board, while we propose **Align**ment **G**enerative **A**dversarial **N**etwork (AlignGAN) to alleviate the remaining displacements.

The contributions of our work is three fold; (1) To the best of our knowledge, this is the first work to generate pixel-level invisibility masks for color images without manual labeling, hence contributing towards the failure detection problem caused by sensory false negatives in autonomous driving. (2) The direct transfer of mid-level representations from color image to thermal image gets promising detection accuracy in thermal domain. (3) Mitigating misalignment problems (AlignGAN) present in color-thermal pairs. Extensive experiments are conducted to quantitatively evaluate the performance of our system.

2 RELATED WORK

Our pixel-level invisibility mask is connected with the task of uncertainty estimation (Gal and Ghahramani, 2016) (Gal, 2016) (Blundell et al., 2015) (Kendall et al., 2015) (Kendall and Gal, 2017), failure detection (Daftry et al., 2016) (Zhang et al., 2014) (Hecker et al., 2018) (Ramanagopal et al., 2018) (Kuhn et al.,

2020) (Corbière et al., 2019) and out-of-distribution detection (DeVries and Taylor, 2018) (Hendrycks and Gimpel, 2016) (Liang et al., 2017) (Lee et al., 2017). Most reported works either consider raw sensor data as introspective inspection (Daftry et al., 2016), estimate the uncertainty of model-based classifiers or compute a confidence score or a binary decision. Our system differs from them by (1) Estimate the confidence of the sensor itself from an outsider's viewpoint, which is the thermal camera; (2) Predicting a confidence probability for every pixel in the image.

Pixel objectiveness (Jain et al., 2017) is the first work to compute pixel-level masks for all *object-like* regions. (Rahman et al., 2019) proposes a failure detection system for traffic signs where excited regions are extracted from feature maps in object detectors to narrow down both the manual labeling space and searching space for false negatives. Though they still need to label the excited regions as false negatives or true negatives. Our work also predicts pixel-level masks for all regions of potentially invisible objects (false negatives) in color images. In contrast, our method doesn't require any manual labeling of the invisible regions, and instead color-thermal image pairs are utilized to provide such supervision.

In order to get aligned color thermal image pairs, (Hwang et al., 2015) created *KAIST Multispectral Pedestrian Dataset* based on a beam splitter to split a beam of light in two for color and thermal cameras. Though image pairs in the data set were observed to have distinct displacements (Zhang et al., 2019) (Li et al., 2018) and also were affected more during night because of the intensity decay of the beam splitter. (Choi et al., 2018) reports less severe disparity problems, though their data sets are not released yet. We observe the same problem in our data set which we collected using the setup of vertically aligned cameras and propose AlignGAN to mitigate such displacements in image planes. The most relevant work to ours is performed at the same time independently by (Vertens et al., 2020). They made similar attempts to create color and thermal image pairs data set called Freiburg Thermal dataset and apply cross modality distillation from color to thermal domain. Though our data set are mainly created for estimating the sensor limitation of color cameras and thus containing a lot of invisible objects in color domain, while theirs targets at the task of semantic segmentation on general nighttime image and their work needs to have objects that are at least visible in both color and thermal domains.

Pix2pix (Isola et al., 2017) and CycleGan (Zhu et al., 2017) developed methods for cross domain translation in paired and unpaired settings. Cycle consistency (Zhu et al., 2017) is the main technique to address unpaired cross-domain translation. Though these translations mainly address style transfers from source to target domains without moving the pixels in the source domain. Our AlignGAN module uses edge maps learned from the thermal image to generate aligned color images.

The remainder of this paper describes our method and experiments in detail. In Section 3, we present our system and describe the network architectures along with the training procedure to generate pixel-level invisibility maps. Finally in Section 4 we conclude with extensive experiments on our data set and show several comparison results. Code, data, and models will be released.

3 SYSTEM OVERVIEW

Our system learns to generate pixel-level invisibility maps for color images in an unsupervised way. During the training phase, the system takes weakly-aligned (color, thermal) image pairs of the same scene as input. Such not-so perfectly aligned image pairs are first registered to remove the geometric differences and then aligned by AlignGAN to remove the remaining displacements between two modalities as detailed in Section 3.1. After the image pairs are well-aligned, our Knowledge *Transfer System* moves the learned representations from the color domain to the thermal domain. Then at the test stage, the pairs are compared to estimate the invisibility score of every pixel in the color image. As a side product, the learned representations of the thermal images can be directly used to construct an object detector for thermal images without any manual labelling or retraining, as shown in Figure 2.

3.1 Alignment Generative Adversarial Network

AlignGAN: From the given poorly aligned raw image pairs, we remove the internal and external transformations between two cameras using standard camera calibration technique, described in Section 4. Then we use AlignGAN to learn how to generate well-aligned color image pairs from weakly-aligned color-thermal pairs. The base block of the systems is shown in Figure 3. As the figure shows, we use two streams for learning - both using the same alignment block - during one iteration of the training phase. The first stream uses a color image and a weakly-aligned thermal image as the edge map computed

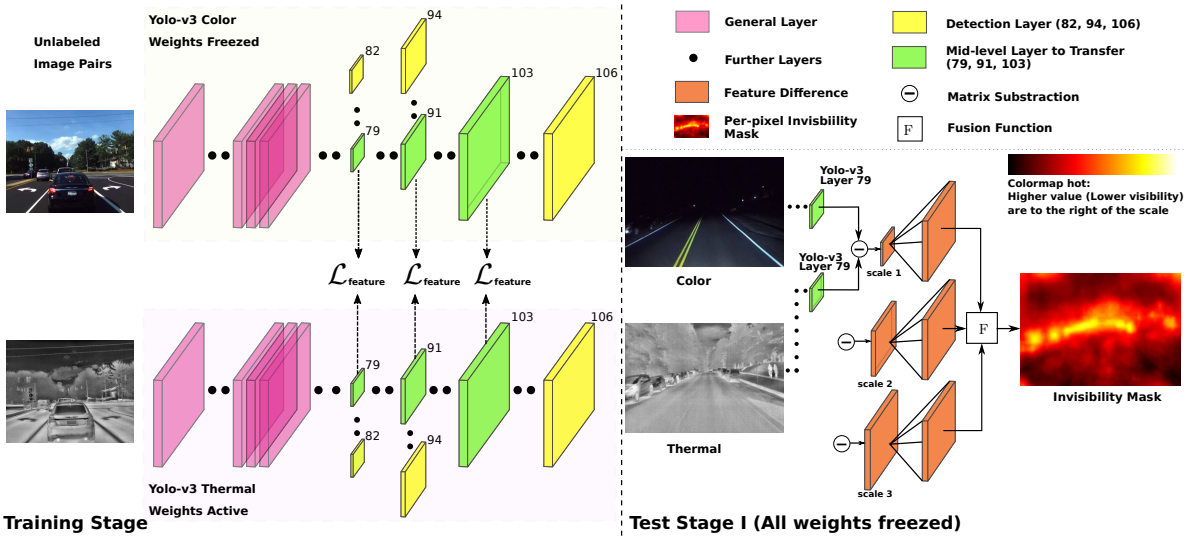

Figure 2: System overview. Our system takes well-aligned image pairs from unlabeled video streams. In the training stage (left), the Yolo-v3 color network is frozen, while Yolov3-thermal network is trained to reproduce the mid-level features as the ones in Yolov-v3 color network at three scales in layer 79, 91 and 103 of the network. In the test stage (right), the difference of those features from layer 79, 91 and 103 are computed and fused to get the final invisibility mask.

based on Canny edge detector as the source and produces a color image as the target which is created using Flownet2 (Reda et al., 2017). Given two consecutive color images from video sequence and one paired image, Flownet2 is used first to estimate the optical flow from one color image to the other. Then we apply Canny edge detector on both the color images and thermal image. The scale of the optical flow is sought within a discrete set of numbers ranging from -0.5 to 1.5 to maximize the overlapping area of edges between source color image of the edge and thermal image of the edge by applying optical flow. In the other stream, source is image is still a color image, edge map though is from a close color image I_{c1} in the video stream, and the target image is the color image I_{c1} itself. We built our system based Pix2pix (Isola et al., 2017). Both the generative network G and G_m use the U-Net architecture (Ronneberger et al., 2015) with an input size of 512×512.

3.2 Knowledge Transfer from Color to Thermal Domains

We choose YOLO-V3 architecture (Redmon and Farhadi, 2018) to run the experiment mainly because (1) It is fast and deployable in real-time applications. (2) No need for a proposal network such as in the Faster RCNN and MASK-RCNN. (3) It has three detection modules based on three different image scales with good detection capabilities for small objects that may appear on a road.

We transfer the learned mid-level features in the

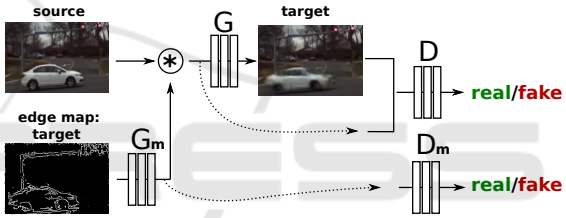

Figure 3: Base alignment network. The module takes as input source image and edge map from target position, and outputs the target image in the given position. The conversion from source to target, performed using network G is conditioned on the motion cues generated using network G_m. The target prediction after network G and motion cues generated by G_m are fed into two discriminators D and D_m respectively.

Yolo-V3 architecture that occurs prior to the detection stage, referred to as the *mid-only* transfer. They are the outputs of layer 79, 91 and 103, respectively from three different scales, as shown in the left side of Figure 2. Similar to (Gupta et al., 2016), we conducted two comparative experiments by transferring the last-layer detection results (Yolo-only) and both intermediate-level features and the detection results (Mid+Yolo). The detection results are computed based on the outputs of detection layers which are 82, 94 and 106.

We construct a thermal detector without any manual labeling and fine-tuning by concatenating learned intermediate features with the detection module from color detector, which produces promising detection accuracies for thermal images in different lighting conditions, as shown in Section 4.2.

3.3 Invisibility Estimation

Now we have two detectors $Yolo_c$ and $Yolo_t$ where $Yolo_c$ is pre-trained on richly-annotated data sets and $Yolo_t$ is trained using well-aligned image pairs to reproduce the same intermediate-level features as the ones in $Yolo_c$. Based on the observation that mid-level features for thermal images are much less affected by the lighting conditions than the ones in color images, as shown in Section 4.2, we use the feature differences to estimate the lighting conditions and thereby estimated the visibility of every pixels in color images.

The YOLO-v3 architecture has three detection modules to predict the objects at three different scales, and consequently it provides intermediate-level features at three different scales. We propose an invisibility score to integrate the features differences at different scales as shown on the right side of the Figure 2. Here we define the invisibility score for an pixel s_i as a function F of the L2-distances of the mid-level features $\{d_k \mid k = 1,2,3\}$ between color and thermal images in Equation 1. Here d_1, d_2, d_3 are from layer 79, 91, 103 respectively and t_k is the highest value that we choose for d_k. Finally we trained a convolutional neural network based on U-Net (Ronneberger et al., 2015) to generate invisibility masks even in the absence of thermal images.

$$F(d_1,d_2,d_3) = 1 - \frac{1}{3}\sum_{k=1}^{3} min((t_k - d_k)/t_k, 0) \quad (1)$$

4 EXPERIMENTS

Experimental Setup: This section presents experimental outcomes for predicting undetectable areas in color images and the unsupervised knowledge transfer from color to thermal domain. We built a sensor platform that can be installed on the roof rack of any car (and we used such as setup for experiments). We used a FLIR ADK camera and the right ZED camera as our sensor pair. The color-thermal calibration was performed using Caltech Calibration tools with a planar checkerboard partly made of an aluminium foil. This removes major parts of camera distortion and helps establish a coarse estimation for common field of view and scale differences. With Homographic warping based on pairs of corresponding points from two image planes, the disparity problem in the static scene can be addressed well. Such weakly-aligned pairs of images are then taken as the input of AlignGAN.

Data Set - Color-thermal Pair: We sampled 18,546 color-thermal pairs from the videos (around 12,0000 image pairs) that we collected while driving in the day to construct the training set for transferring the intermediate-level features from color domain to thermal domain. For the validation data set, we manually labeled 2000 image pairs with object bounding boxes, 500 during dawn, 500 during dusk, 500 during night and 500 during fog. They were used to evaluate the prediction performance of the undetectable area in color images and detection performance for thermal images. We don't have the exact statistics of our training set since we didn't label them by hand. Though our manually labeled validation set which contains 7821 cars, 1138 traffic signs, 626 people and 343 trucks provide a sense of our training data set.

Comparison with Other Data Sets: To the best of our knowledge, our data set is the first to attempt on estimating pixel level sensor limitation for color cameras. Because it is based on color-thermal image pairs, even objects that are not seen from the color domain will also be annotated. The Freiburg Thermal Dataset (Vertens et al., 2020) that consists of (color, thermal) image pairs could be very suitable for evaluating our invisibility score, although it has not been released yet. Other data sets like Dark Zurich (Sakaridis et al., 2019) or Foggy Zurich Data sets (Dai et al., 2020) may not be as suitable because of the following two reasons, even though both of them are captured under challenging visibility conditions. First, since both of them only have images from the color domain, it's impossible to annotate completely dark/heavy foggy areas where objects are actually present. Secondly, both data sets are much less challenging than Color-thermal Pair Data sets. Through close observation, almost all the 2617 night images from Dark Zurich are captured from well-lit city streets, making most objects already visible to the color images, which is shown in the top right plot in Figure 4. Among the 3808 images from the Foggy Zurich data set, nearly half of them are captured in very light fog and the remaining images contain very few objects that are challenging to detect, which is also shown in the bottom right plot in Figure 4. Dark Zurich data set has 50 images that are annotated and we included and expanded the test data set by manually labeling bounding boxes of objects for other 46 images from its test data, having 96 images in total. Foggy Zurich Data has 40 images that are annotated. Similarly we expand the test data set to 80 images by manually annotating 40 more sampled from its data set.

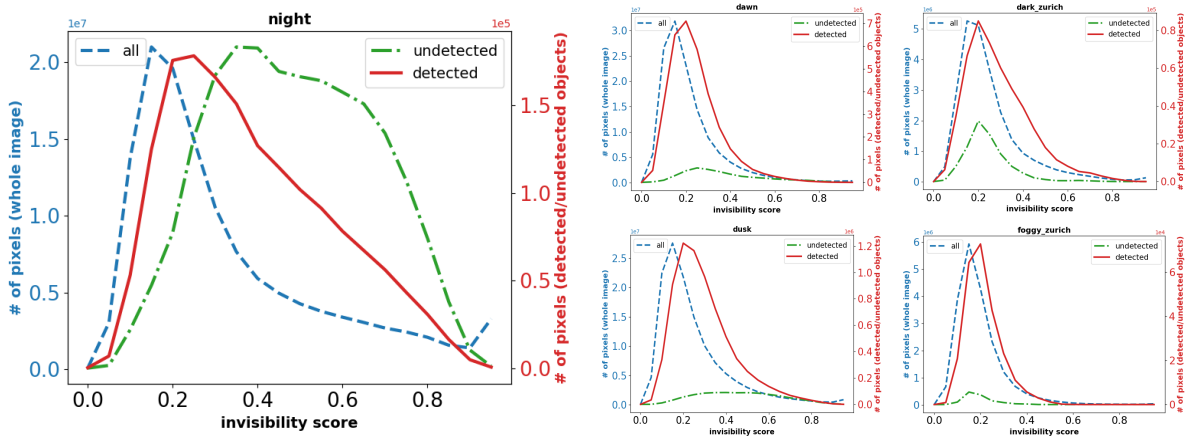

Figure 4: We show the result of pixel-level invisibility masks on capturing false negatives of MASK-CNN for night scenes (left). With a threshold of 0.35 (separated by empty bars and solid bars) for invisibility score, the invisibility map can cover 77.5% of the pixels in the undetected objects(green bar) while only taking up 35.9% of the pixels in the night images. Performance of the invisibility score from dawn, dusk time and other data sets of Dark Zurich and Foggy Zurich are also shown on the right.

Experiment Focus: We focus on answering the following five questions in this section. (1) How good is the prediction of invisible areas in color images? (2) How good is the detection performance on thermal images through knowledge transfer? (3) Which level of representation transfer will give the best result on the detection accuracy? (4) Will our AlignGAN enhance the knowledge transfer process? and finally (5) How will the transfer performance change with respect to the number of images pairs? Now we answer those questions quantitatively using our results.

4.1 How Good is the Prediction of the Undetected Areas in the Color Image?

We use intermediate-level features from two paired DNNs as a shared space where Euclidean distance serves as the estimation of the reliability of color images compared to thermal images. Our experiment results show that the proposed system can produce good quality masks for invisibility. In our experiment, we set the t_1, t_2, t_3 to 4, 3.5, 3.2 respectively in Equation 1.

Our first observation is that the predicted invisibility scores change based on the light intensity variations of the environment. For image pairs respectively from day, dawn, dusk, and night, we compute the invisibility score learned from our system and the L2 distance between the intermediate-level features of color and thermal images. As shown in Figure 5, both the feature difference and invisibility score increases while the light intensity of the environment decreases. This is consistent with the observation that color images are much less reliable than thermal images under poor lighting conditions. The Gaussian distribution of invisibility score for day is (0.020, 0.028) and for night is (0.268, 0.052). The two distributions are highly separable with few overlaps, as shown in Figure 5, which qualitatively demonstrates that our pixel-level invisibility score is able to predict the unseen regions in the color images under varying lighting conditions.

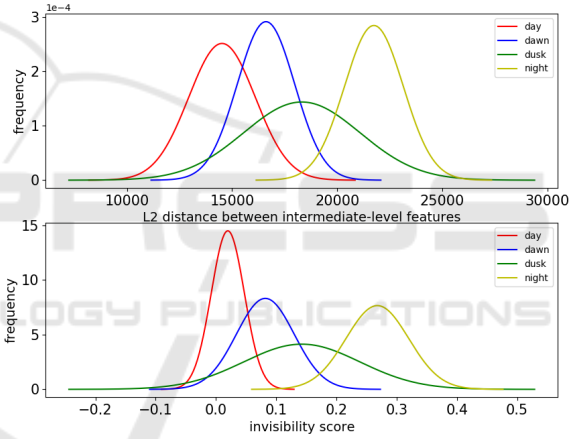

Figure 5: Distributions of distances and invisibility score. We show the L2 distance distributions (upper figure) between intermediate-level features from layer 79, 91 and 103 in Yolo-v3 and the invisibility scores (lower) from different lighting conditions. The Gaussian distributions from left to right are respectively from data of daytime, dawn, dusk and night.

A good binary visibility mask has two characteristics of (1) covering most of the undetectable objects (2) covering only the undetectable area in an image, which we use in quantified form to assess the effec-

Table 1: Detection accuracy for different lighting conditions. Even for night scenes which are not present in the training set, the overall accuracy is up to 34.2%.

Name	Car	Person	Traffic Light	All
day	**0.620**	0.191	**0.576**	**0.462**
dusk	0.610	0.186	0.485	0.424
dawn	0.600	**0.337**	0.476	0.471
night	0.506	0.120	0.399	0.342
fog	0.496	0.149	0.365	0.337

Table 2: Detection accuracy for different layers. We found that using only the mid-level features to transfer can achieve the best accuracy for object detection.

Name	Car	Person	Traffic Light	All
Mid-only	**0.513**	**0.241**	0.470	**0.408**
M+Y 0.05	0.470	0.178	**0.516**	0.388
M+Y 0.1	0.474	0.171	0.500	0.382
M+Y 1.0	0.437	0.170	0.475	0.361
Yolo-only	0.447	0.164	0.463	0.358

tiveness of our results.

Now we report results of a quantitative analysis showing the effectiveness of predicting the undetectable area in the color images. The invisibility mask for nighttime images can cover 77.5% of the undetectable area in the image with the visibility threshold of 0.35 and only report 35.9% of the whole image. The invisibility mask for dawn time covers 49.2% of the undetected area while reports 15.4% of the whole image. For dusk, it covers 73.4% of the undetected area while reports only 22.2% of the image, as shown in Figure 4.

We also tested our invisibility estimation on model failures by evaluating our color-only model on two expanded data sets from Dark Zurich and Foggy Zurich. The number of undetected objects are far less than the detected objects compared to our color-thermal data set, which is shown in the right most plots of Figure 4. Furthermore, the invisibility map can cover only 15% of the undetected objects from MASK-RCNN. This justifies the claim that sensor limitation is different from model limitation and our invisibility score is a good indicator of the limitation of the color camera itself.

4.2 Can Paired Data Facilitate Detection using Transfer Learning?

Firstly, we report that the knowledge transfer through mid-level features can reach 46.2% overall detection accuracy for thermal imagery in the Color-thermal data set. Because learning doesn't require any manual annotations and doesn't require any retraining, we found the result to be promising.

We evaluated the effectiveness of AlignGAN using the post application of object detection. We tested the detection performance of the daytime data trained thermal detector on the night time data, and observed that it can still get to an overall accuracy of 34.2%. The images used in the training phase were chosen from day time with good lighting conditions and the test set includes images taken during dawn, dusk, night and fog. Without any manually labelled training data, the detection IOU of cars can reach 50.6% in the night as shown in Table 1. This quantity shows that the intermediate-level features of thermal images can be transferred smoothly from day to night, in contrast to the ones in color images. The principal used in our invisibility score is that *features in thermal images remain stable with respect to light change and are trained to be like the ones in color, the mid-level features produced by thermal images are of the same utility of features in color images when lighting conditions are poor*. This is the most possible reason for the success on estimating the invisibility of the color images using our invisibility scores as shown in Section 4.1.

4.3 Which Layer is More Effective?

We experimented with the knowledge transfer from different layers. Gupta et al. (Gupta et al., 2016) indicates that combining mid-level features with last layer features will give the best detection results when retraining on the target data sets. Although we show that using mid-level features only gives the best detection result of (40.8%) over mid-last layer transfer (36.1%) and the one using the yolo-layer only gives (35.8%). These results are summarized in Table 2. More importantly, we varied the weight of the Yolo layer to 0.05 and 0.1 in the loss function and conducted two more experiments. Surprisingly, we found that higher weights on the Yolo layer resulted in worse overall detection accuracy. One potential explanation for this observation is that the data set is not large enough to train a new object detector for all modules, especially for both the class prediction and bounding box regression. Consequently, we learnt that it is more efficient to learn only the mid-level features and to not change layers of the detection module.

Table 3: Detection accuracy for AlignGAN. With the alignment module, there is a relative 2.77% boost in terms of overall accuracy.

Name	car	Person	traffic light	All
Mid+Yolo 1.0	0.437	0.170	0.475	0.361
Mid+Yolo 1.0 + Flow Encoder	**0.442**	**0.191**	**0.481**	**0.371**

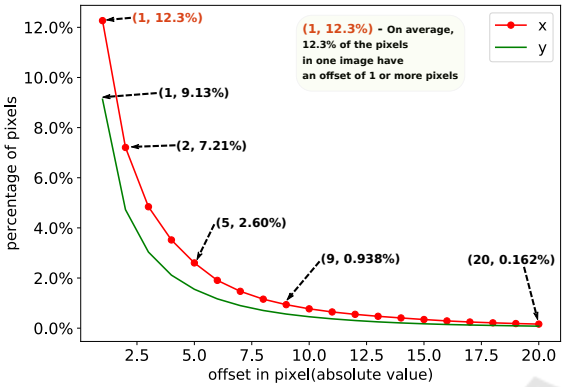

Figure 6: X-Y offset distribution. We show that in the image plane offset in x direction is on average larger than that in y direction. Also distinct movements(5 pixels or more) constitute 2.6% of the pixels in the images.

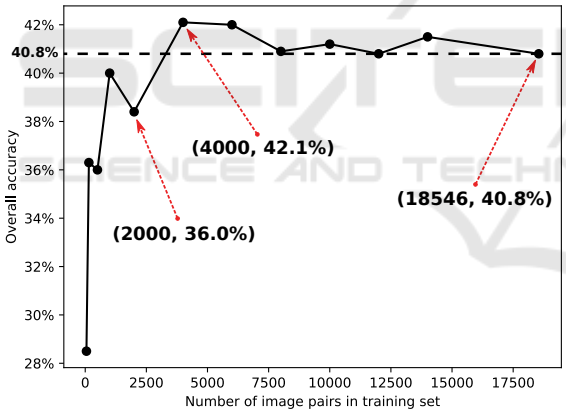

Figure 7: Comparison of data set sizes. We experiment with different sizes of training data sets for knowledge transfer, and observes that using 4000 image pairs can obtain the same if not better accuracy as the original size of 18,546 image pairs.

4.4 How Much will the AlignGAN Help the Detection?

Even after the pre-processing, the image pairs still have some displacements. Here we calculate the statistics of such displacements. With an image resolution of 640 by 512, the estimated X-Y displacement is shown in Figure 6. On average, there are only about 2.6% pixels in one image that have 5 or more pixel displacements on the X-direction, which we consider to be the threshold of movement that affects the detection results. Also, we noticed more displacements in the X-direction than in the Y-direction in the image plane. We attribute this to the fact that object movements projected to the image plane are more obvious in X-direction than the Y-direction.

We evaluated how AlignGAN can improve the data transfer from color domain to thermal domain. In Table 3, the Alignment module gave an enhancement of 1% in terms of detection accuracy.

4.5 How Many Pairs are Needed to Get a Good Transfer?

Section 3.2 showed that mid-level feature transfer results in the best performance. We now determine the number of image pairs required to achieve that performance. Surprisingly, Figure 7 shows that randomly sampled 4000 image pairs from the space of 18,546 image pairs can achieve the same if not better accuracy than the entire sample space. This observation implies that the domain difference from color to thermal can be learned from a small amount of images pairs and the transfer discriminating visual representations from the well-established color detection task to thermal images can be done in a light-weighted manner.

Figure 7 shows experimental results with different number images from 100 to 18000 and that the performance will be stable after 4000 images. Such results may appear to be counter intuitive at first sight as more data often leads to better results when model capacity is large enough like the one we use in the experiment. One potential explanation for the saturation point is that the two domains have much in common and thus the domain difference can be mitigated with a few examples. This saturation point observation with 4000 samples can be used as a promising deployment strategy of direct knowledge transfer. One conclusion from this observation is that less time is needed for training for both the invisibility prediction and the direct knowledge transfer for object detection.

5 CONCLUSION

Given a color image, we predict a pixel-level invisibility mask for such an image without manual labelling. Equipped with this map, a system could decide to trust detection results of some regions over others in an image or generate warnings to a driving automation. Such a mask can also be used as the confidence map to fuse the outputs from multiple sources when many visual sensors are deployed. Compared to existing works in the area, our mask estimates the sensor limitations for each pixel instead of model limitations, which is the first attempt to fill this gap. Experiments have shown that our mask is able to distinguish most of the invisible pixels from the visible pixels. Our results also demonstrate the effectiveness of building an object detector for the thermal domain using the mid-level features transferred from its peer color images. Our ongoing work is on creating a generalized invisibility mask for color, thermal, Lidar and radar imagery. And more importantly we will explore an approach to detect scenarios when all the sensors will fail, so that a driving automation could invoke appropriate failure tolerance mechanisms.

ACKNOWLEDGEMENTS

This research was partially supported by the Commonwealth Cyber Initiative, which invests in cyber R&D, innovation, and workforce development (cyberinitiative.org). Some of the experiments were run on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University, VA. (URL: http://orc.gmu.edu). We also thank Meili Liu for labeling the validation data set for the experiments.

REFERENCES

Aytar, Y., Vondrick, C., and Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, pages 892–900.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*.

Choi, Y., Kim, N., Hwang, S., Park, K., Yoon, J. S., An, K., and Kweon, I. S. (2018). Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948.

Corbière, C., Thome, N., Bar-Hen, A., Cord, M., and Pérez, P. (2019). Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, pages 2898–2909.

Daftry, S., Zeng, S., Bagnell, J. A., and Hebert, M. (2016). Introspective perception: Learning to predict failures in vision systems. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1743–1750. IEEE.

Dai, D., Sakaridis, C., Hecker, S., and Van Gool, L. (2020). Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. *International Journal of Computer Vision*, 128(5):1182–1204.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

DeVries, T. and Taylor, G. W. (2018). Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338.

Gal, Y. (2016). Uncertainty in deep learning. *University of Cambridge*, 1:3.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Gupta, S., Hoffman, J., and Malik, J. (2016). Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Hecker, S., Dai, D., and Van Gool, L. (2018). Failure prediction for autonomous driving. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1792–1799. IEEE.

Hendrycks, D. and Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Hwang, S., Park, J., Kim, N., Choi, Y., and So Kweon, I. (2015). Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference*

on computer vision and pattern recognition, pages 1125–1134.

Jain, S. D., Xiong, B., and Grauman, K. (2017). Pixel objectness. *arXiv preprint arXiv:1701.05349*.

Kendall, A., Badrinarayanan, V., and Cipolla, R. (2015). Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*.

Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584.

Kuhn, C., Hofbauer, M., Petrovic, G., and Steinbach, E. (2020). Introspective black box failure prediction for autonomous driving. In *31st IEEE Intelligent Vehicles Symposium*.

Lee, K., Lee, H., Lee, K., and Shin, J. (2017). Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*.

Li, C., Song, D., Tong, R., and Tang, M. (2018). Multispectral pedestrian detection via simultaneous detection and segmentation. *arXiv preprint arXiv:1808.04818*.

Liang, S., Li, Y., and Srikant, R. (2017). Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer.

Rahman, Q. M., Sünderhauf, N., and Dayoub, F. (2019). Did you miss the sign? a false negative alarm system for traffic sign detectors. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3748–3753.

Ramanagopal, M. S., Anderson, C., Vasudevan, R., and Johnson-Roberson, M. (2018). Failing to learn: autonomously identifying perception failures for self-driving cars. *IEEE Robotics and Automation Letters*, 3(4):3860–3867.

Reda, F., Pottorff, R., Barker, J., and Catanzaro, B. (2017). flownet2-pytorch: Pytorch implementation of flownet 2.0: Evolution of optical flow estimation with deep networks.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Redmon, J. and Farhadi, A. (2017). Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Sakaridis, C., Dai, D., and Gool, L. V. (2019). Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7374–7383.

Vertens, J., Zürn, J., and Burgard, W. (2020). Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. *arXiv preprint arXiv:2003.04645*.

Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., and Darrell, T. (2018). Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*.

Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z., and Liu, Z. (2019). Weakly aligned cross-modal learning for multispectral pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5127–5137.

Zhang, P., Wang, J., Farhadi, A., Hebert, M., and Parikh, D. (2014). Predicting failures of vision systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3566–3573.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.