Multispectral Semantic Segmentation for Autonomous Vehicles: Optimizing Detector Weights As A Function of Daytime and Visibility

Francesco Mastrocinque
Departments of Chemistry & Electrical and Computer Engineering
Duke University
Durham, NC 27701
francesco.mastrocinque@duke.edu

Abstract

This work addresses the semantic segmentation of images of street scenes for autonomous vehicles using a RGB-Thermal dataset. An increasing interest in safe and functioning autonomous vehicles has lead to the adoption of image segmentation in this area. Research relating to autonomous vehicle semantic segmentation and object detection has traditionally relied on the use of RGB images acquired during times of poor visibility or adverse weather conditions, or through the use of expensive light detection and ranging (LIDAR) technology. This work focuses on the application of imaging techniques to improve the current state-of-the-art image segmentation models for autonomous vehicles by incorporating information obtained from both RGB and thermal-based input images. The results described herein show a drastic improvement in model image segmentation as identified through a higher average Sorensen-Dice coefficient when both RGB and thermal information were used across all times of day, as opposed to models trained using RGB or thermal-only channels.

1 Introduction

Autonomous vehicles have increasingly drawn attention from the vehicle and transportation industry, and are expected to change the future of transportation dramatically.[1] To carry out the successful navigation of autonomous vehicles requires the combination of robust technologies that spans across many different disciplines such as electrical engineering, mechanical engineering, computer science.[2] As a system that undergoes autonomous transport, these vehicles must have a method for sensing their environments to make calculated decisions such as lane changing, accident avoidance, and obeying traffic signs. However, despite autonomous vehicles becoming an increasingly important area, the related technology is far from mature.[3]

Autonomous vehicles have to be perceptive of their surroundings in real-time before they are able to make an actionable decision. Current perception technologies relevant to autonomous vehicles involve costly instruments, such as ultrasound, radar, and laser sensors.[4] Image segmentation involving inexpensive alternatives, such as using cameras, have been previously explored but are typically based on visible (RGB) images only.[5] However, RGB-only images are impractical during times of poor visibility, such as at night or in adverse weather conditions. In order for autonomous vehicles to not only be robust and practical, they should be able to navigate safely, independent of the environmental visibility conditions. Therefore, building a model capable of performing semantic segmentation using only RGB-based images is not sufficient. Instead, models capable of semantic segmentation for autonomous vehicles should utilize multispectral input images to better interpret vital contextual information of the vehicle's surroundings. This work explores the impact

of multispectral information (RGB and thermal infrared) on semantic segmentation as a function of daytime and visibility conditions.

2 Related Work

Deep learning convolutional neural network (CNN) models have proven successful in several computer vision tasks, such as in image classification,[6] segmentation,[7] and object detection.[8] Regarding semantic segmentation, in 2015, Long et al. proposed a segmentation network they coined as Fully Convolutional Networks (FCNs) for segmenting a number of visual object classes in realistic scenes.[9] Shortly thereafter, the concept of an encoder/decoder network architecture for semantic segmentation was introduced in SegNet, which aimed to increase model performance with respect to image segmentation. [10] The SegNet network featuring an encoder/decoder architecture increased model performance through the use of several encoding and decoding blocks that are responsible for downsampling and upsampling the image, respectively, while preserving high resolution information at each layer. These convolutional neural networks provided a solid foundation for the genesis of several rapid deep learning architectures for computational semantic segmentation tasks.

Regarding semantic segmentation of multispectral images relevant to autonomous navigation, Ha et al. created a new encoder/decoder architecture involving two distinct encoder blocks for each of the input multispectral channels (one for RGB images, and the other for thermal images) that were combined before entering the appropriate decoding block.[3] Although this network performed much faster than SegNet, with an inference time of almost an order of magnitude smaller, the network was not robust enough to achieve a high Jaccard similarity index (0.64). Instead, Unet-type deep learning architectures have repeatedly been shown to be robust enough to semantically segment images across a diverse range of inputs with relatively rapid inference times while maintaining high Jaccard similarity indices.[11]

Here, I propose utilizing a custom Unet architecture along with a physical layer capable of optimizing multispectral detector weights as a function of daytime and visibility. This approach exploits the strengths and intrinsic contextual information embedded in both of the RGB and thermal channels, while making critical decisions about what contextual information is crucial for model performance as a function of daytime and visibility. Such a robust deep learning network could be capable of assessing environmental and visibility conditions to enable self-navigation of autonomous vehicles via appropriate semantic segmentation in real time.

3 Methods

3.1 Dataset

The dataset was obtained from the following website: www.mi.t.u-tokyo.ac.jp. Data consisted of 1,569 mixed images (820 taken at daytime and 749 taken at nighttime). Each image file terminated with either a letter "D' or "N" indicating if the images were taken during the daytime or nighttime, respectively. These RGB and thermal images were captured using an InfRec R500 as the camera. Images were taken using a spatial resolution of 480x640 pixels. Eight classes of obstacles commonly encountered during driving (cars, pedestrians, bicycles, curves, crosswalks, guardrails, traffic cones, and speed bumps) were labeled in the appropriate masks of the respective input images. Label images were provided as 480x640 pixel masks containing a single channel, with each object class having a different pixel value (1 for cars, 2 for pedestrians, etc.). The dataset was randomly divided into training and test data using an 80/20 training/test split.

3.2 Pre-processing and Network Architecture

A new mask object was created that reshaped, transformed, and stored each object class in the original mask image to occupy its own channel. For example, objects that contained a pixel value of 1 (cars) were placed into the first channel, whereas objects that contained a pixel value of 2 (pedestrians) were placed in the second channel, and so on and so forth. Both the input images and masks were resized to 128x128 pixels prior to model training in order to reduce computational load.

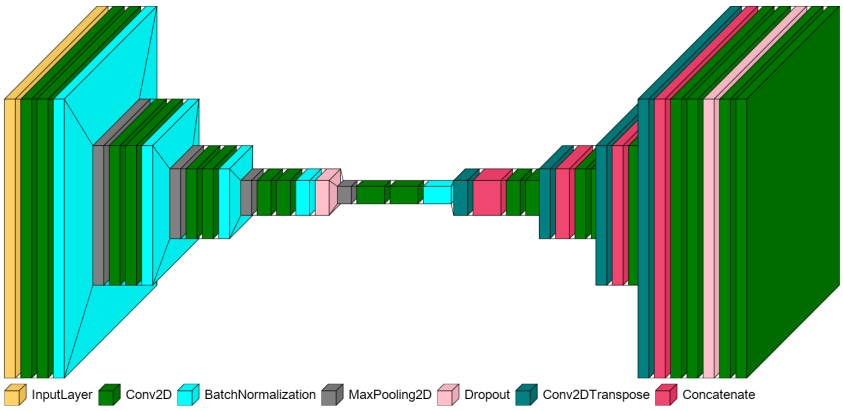

Figure 1: Representative Unet Architecture used in this study. Physical layer not shown for conceptual clarity.

In order to assess whether or not multispectral images can increase model robustness for semantic segmentation throughout varying times of visibility, models trained using either RGB or thermal-only channels were used as baseline models. Furthermore, to assess whether or not weighted multispectral inputs of RGB or thermal channels can provide better model segmentation throughout various times of day, as opposed to multispectral images that utilize an unweighted distribution of all channels, unweighted multispectral images were also used as a baseline. Each baseline model was trained in three distinct conditions:

1. Using daytime images only
2. Using nighttime images only
3. Using all images

This sets a fair benchmark to assess whether a physical layer is needed to effectively segment images at a higher performance as compared to single spectra (RGB or thermal) or unweighted multispectral input images with respect to image visibility (daytime vs. nighttime). In summary, 12 deep learning networks were trained in this study (RGB-only, thermal-only, unweighted multispectral, and weighted multispectral, each trained exclusively on images taken during the daytime, nighttime, and when all images are used).

A custom Unet was designed and implemented in this study (**Figure 1**), which featured four encoding and four decoding blocks. Each encoding block consists of two convolutional layers, one batch normalization layer, and a maxpooling layer. Prior to entering the bottleneck block, a dropout layer was added to prevent the model from overfitting the training data. Each decoding block contains a 2D convolutional transpose layer responsible for upsampling, 2 convolutional layers, and a concatenation layer that functioned as the skip connection. After the final decoding block, another dropout layer was added, followed by two terminal convolutional layers.

The physical layer was implemented only for the architecture involving the trainable weight parameters for the weighted multispectral model. The physical layer was implemented directly after the input layer, prior to entering the Unet. The physical layer was constrained to produce trainable weights that were not only positive in value, but summed to a value of one, in order to express the dependence on either the RGB channels or the thermal channels as a percentage for a given time of day. The weights were initialized at a value of one for both the RGB and thermal channels before training. Before passing the multispectral images into the physical layer, the RGB channels were averaged to produce an image object that was reduced from 4 channels (R, G, B, Thermal) to 2 channels (RGB average, Thermal).

3.3 Training

The model architecture described above was implemented using Tensorflow and trained on an NVIDIA GeForce RTX 2070 Super GPU. All of the models were trained for 100 epochs, with an initial learning rate of 0.001. Custom callback functions were implemented and used to decrease learning rate an order of magnitude as the model loss started to converge (see code files).

3.4 Performance Metrics

Model performance was evaluated using the average Sorensen-Dice coefficient of the test dataset. The Sorensen-Dice coefficient can be described mathematically as:

$$SDC = \frac{2*|X \cap Y|}{|X \cup Y|}$$

The Sorensen-Dice coefficient is a value that is fixed between a low of zero and a high of one. It is a measurement that describes the degree of overlap between a model's predicted mask and ground truth mask.

4 Results

Table 1: Average Sorensen-Dice coefficients

Simulation	Daytime	Nighttime	All
RGB-Only	0.7326	0.5304	0.7156
Thermal-Only	0.7192	0.5786	0.7132
Unweighted Multispectral	0.7664	0.6473	0.7706
Weighted Multispectral	0.8380	0.7137	0.8647

Results are summarized in **Table 1**. Overall, average Sorensen-Dice coefficients evaluated on the test datasets ranged from ~0.53 (RBG-only channel at nighttime) to ~0.86 (weighted multispectral channels using mixed images). Unsurprisingly, the baseline models involving RGB-only and thermal-only channels for input data performed relatively poorly for image segmentation on images collected exclusively at night. Models trained using RGB-only and thermal-only channels also performed fairly well during the daytime or when using all images mixed together, achieving Sorensen-Dice coefficients around ~0.7.

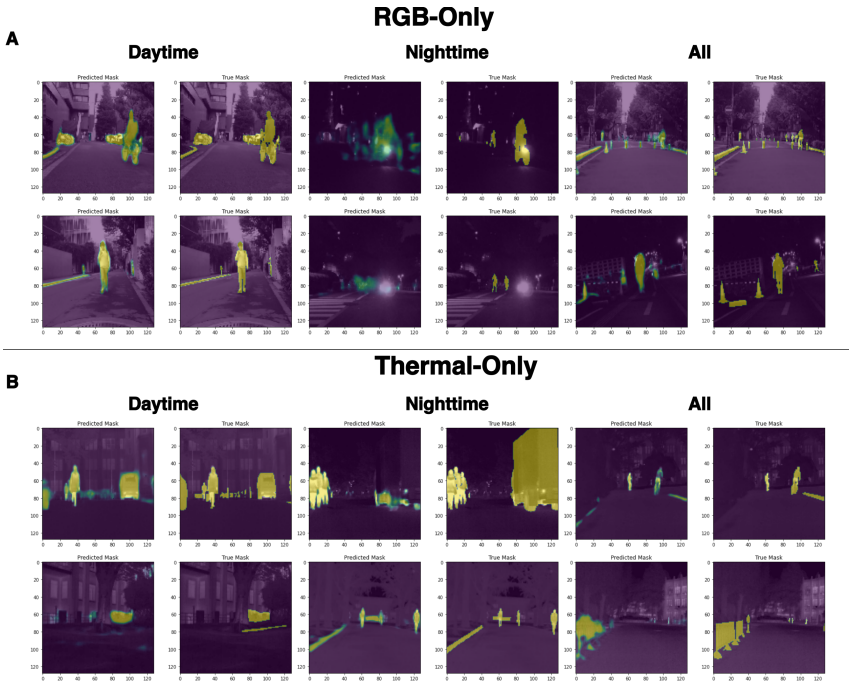

Figure 2: Side by side comparison of model predicted masks by models after training and respective ground truth masks. Both overlayed with original input image. **A)** RGB-only channel model generated masks. **B)** RGB-only channel model generated masks.

Table 2: Physical Layer Trained Weights As A Function of Visibility

	Daytime	Nighttime	All
RGB Channel Weight	0.564	0.312	0.319
Thermal Channel Weight	0.436	0.688	0.681

The model trained using unweighted multispectral data performed significantly better than the models trained using RGB-only or thermal-only channels across all times of day, achieving Sorensen-Dice coefficients that ranged from ~0.64 (night) to ~0.77 (mixed). Upon implementing a physical layer that learned trainable parameters to optimize a weighted distribution of RGB and thermal channels of the input images for image segmentation, the model chose weightings that varied as a function of daytime and visibility (**Table 2.**). The model that incorporated a physical layer significantly outperformed all other models in this study across all times of day.

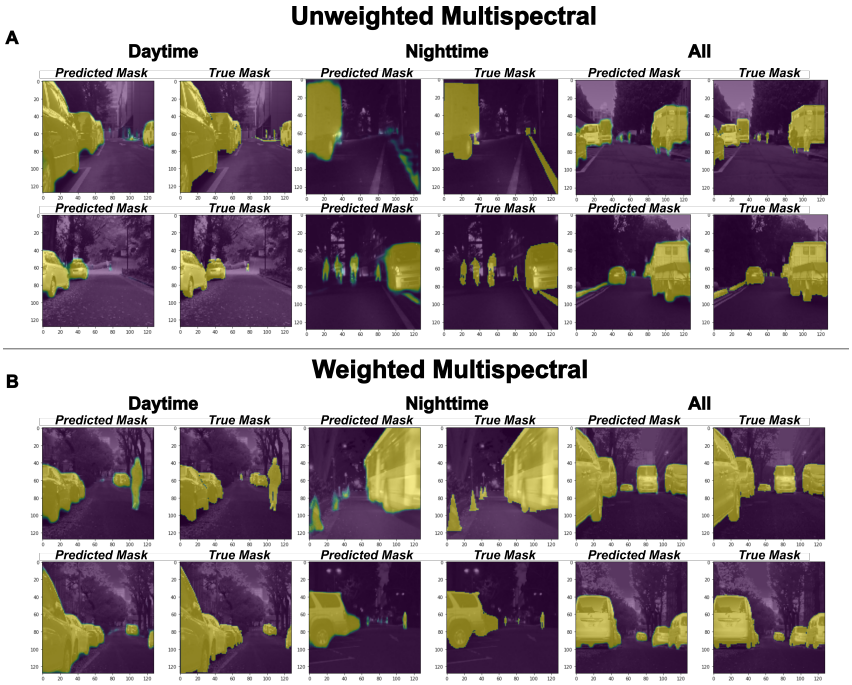

Figure 3: Side by side comparison of model predicted masks by models after training and respective ground truth masks. Both overlayed with original input image. **A)** Unweighted multispectral model generated masks. **B)** Weighted multispectral model generated masks.

5 Discussion

In this work, I designed and implemented a custom Unet model capable of finding an optimal weighting between RGB and thermal channels for semantic segmentation of images relevant to autonomous vehicles and self-navigation. As noted above, the RGB-only and thermal-only baseline models performed relatively poorly on their own in times of poor visibility, characterized by their relatively low Sorensen-Dice coefficient for segmentation of images taken exclusively at night. Specifically, the RGB-only model trained using images taken at night was the worst model explored in this study, yielding ill-defined segmented silhouettes of people and an average Sorensen-Dice coefficient of ~0.53 (**Figure 2A**). On the contrary, the model trained using RGB-only channels on images taken during the daytime performed much better, achieving an average Sorensen-Dice coefficient of ~0.73. **Figure 2A** qualitatively shows that the daytime RGB-only model was able to segment all object classes fairly well. In addition, the RGB-only model performed well when trained on all images (Sorensen-Dice coefficient: ~0.72), likely due to being trained on a larger, more diverse set of images. However, segmentation of images taken at night using the model trained using all images as input training data still performed poorer on average than the daytime images when the

model was evaluated on the test set. This model performed most poorly for image segmentation of objects that were greater distances away from the camera, as these objects are not as well resolved as objects closer to the imaging camera.

Like the RGB-only model, the model trained using only the thermal channel of daytime images were segmented well across almost all object classes (Sorensen-Dice coefficient: ~0.72). However, across all times of day, the thermal models all performed poorly on objects that do not typically give off much heat relative to their surroundings, such as sitting cars, curves that define edges of the road, traffic cones, and metal fences. However, this model was very consistent at segmenting people across all times of day (**Figure 2B**).

The unweighted multispectral models trained during the daytime, nighttime, and with mixed visibility images all produced networks that achieved higher average Sorensen-Dice coefficients than the RGB-only and thermal-only models across all times of day. Clearly, having both the RGB and thermal channels available helped the model obtain information that is absent in just the RGB or thermal-only channels. Thus, thermal and RGB channels both contain important, yet distinct information that is key to image segmentation, independent of the time of day. **Figure 3A** shows representative predicted masks made by the unweighted multispectral model.

Lastly, when the model was provided a physical layer to optimize the image weights prior to entering the Unet architecture, the model chose different RGB and thermal weightings depending on the time of day (**Figure 3B**; **Table 2**). During the daytime, the weighted model found optimal input image weights that relied on the RGB and thermal channels almost equally (~0.56 and ~0.44, respectively). This is likely because the thermal channel both perform almost equally well at segmenting images that were taken during the day. As mentioned previously, the thermal camera does an excellent job at providing distinct images of pedestrians, whereas the RGB camera is able to clearly pick up information about vehicle and object outlines. The network found that weighting on the RGB and thermal channels equally during the day gave the most optimal performance for image segmentation At night and when all images were used as inputs, the model placed more weight on the thermal channel of an input image rather than the RGB channel (~0.69 and ~0.31, respectively). Again, this is likely due to the fact that the thermal camera is able to produce well-defined silhouettes of pedestrians and active vehicles as these objects give off much more heat than their surroundings. At night, the RGB camera is only able to pick up outlines of objects that are relatively close to the imaging device. Finally, the model found an optimal weighting that again relied more heavily on the thermal channel than the RGB channel when trained using all images. This is likely because half of the images are taken at night, and the thermal channel can be utilized to produce well-segmented images both during the day and at night, as opposed to the RGB channels which tend to only perform well on images taken during the day.

In summary, it's clear that RGB and thermal images both contain important contextual information useful for image segmentation of images taken either at night or during the daytime. Models that took advantage of all RGB and thermal information (such as the weighted and unweighted multispectral models) outperformed the models that utilized exclusively one or the other during both the daytime and nighttime. When a physical layer was implemented, the model determined that it was most effective to have a near equal weighting between the RGB and thermal channels to attain an optimized input image for segmentation for images taken during the day. During times of poor visibility (night), your autonomous vehicle should place more weighting on images rendered through the thermal camera rather than the RGB camera.

Acknowledgements

I would like to acknowledge Dr. Horstmeyer, Kanghyun Kim, and Shiqi Xu for their useful suggestions, guidance on this project, and assistance throughout the semester.

References

[1] Rasheed Hussain and Sherali Zeadally. "Autonomous Cars: Research Results, Issues, and Future Challenges". In: *IEEE Communications Surveys Tutorials* 21.2 (2019), pp. 1275–1313. DOI: 10.1109/COMST.2018.2869360.

[2] Keshav Bimbraw. "Autonomous cars: Past, present and future a review of the developments in the last century, the present scenario and the expected future of autonomous vehicle technology". In: 01 (2015), pp. 191–198.

[3] Qishen Ha et al. "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes". In: (2017), pp. 5108–5115. DOI: 10.1109/IROS.2017.8206396.

[4] X. Chen et al. "Multi-view 3D Object Detection Network for Autonomous Driving". In: (2017), pp. 6526–6534. ISSN: 1063-6919. DOI: 10.1109/CVPR.2017.691. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.691.

[5] Marius Cordts et al. "The Cityscapes Dataset for Semantic Urban Scene Understanding". In: *CoRR* abs/1604.01685 (2016). arXiv: 1604.01685. URL: http://arxiv.org/abs/1604.01685.

[6] G. Chen et al. "Tensor Network for Image Classification". In: (2020), pp. 135–140. DOI: 10.1109/ICDH51081.2020.00031. URL: https://doi.ieeecomputersociety.org/10.1109/ICDH51081.2020.00031.

[7] Fabian Isensee et al. "NNU-net: A self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* (2020). URL: https://www.nature.com/articles/s41592-020-01008-z#citeas.

[8] Khuram Faraz et al. "Deep learning detection of nanoparticles and multiple object tracking of their dynamic evolution during in situ Etem Studies". In: *Scientific Reports* (2022). URL: https://www.nature.com/articles/s41598-022-06308-2.

[9] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully Convolutional Networks for Semantic Segmentation". In: *CoRR* abs/1411.4038 (2014). arXiv: 1411.4038. URL: http://arxiv.org/abs/1411.4038.

[10] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation". In: *CoRR* abs/1511.00561 (2015). arXiv: 1511.00561. URL: http://arxiv.org/abs/1511.00561.

[11] Nahian Siddique et al. "U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications". In: *IEEE Access* 9 (2021), pp. 82031–82057. DOI: 10.1109/ACCESS.2021.3086020.