# Adversarial Multi-Modal Transfer Learning

Connor Roos
University of Virginia
Charlottesville, Virginia
cgr3mu@virginia.edu

Kishan Athrey
University of Central Florida
Orlando, Florida
kishan.athrey@knights.ucf.edu

Mubarak Shah
University of Central Florida
Orlando, Florida
shah@crcv.ucf.edu

## Abstract

*This Summer research concerns utilizing Adversarial Network Compression in order to improve the accuracy of visual pedestrian detection neural networks in adverse lighting conditions. The bulk of this research was accomplished during the 2018 Summer Research Experience for Undergraduates held Center for Research in Computer Vision at the University of Central Florida. This research was supported by NSF Grant CNS-1757858.*

## 1. Introduction

Currently some state-of-the-art Pedestrian Detection models utilize both visual and thermal data in order to improve precision and lower miss rate by simply adding thermal data as another mode[5]. It should be noted however that the majority of surveillance systems employed for such a task solely use visual sensors and thus lack the improved accuracy that can be gained by utilizing thermal data.

This paper introduces a novel way of improving visual-only neural networks by transferring the implicit relations between visual and thermal data by transferring the implicit relations used to obtain better quantitative results in a visual-thermal "teacher" model to a smaller visual-only "student" model. The method described in this paper can be applied to any number of different input modes and can be utilized with any combination of student and teacher models, regardless of the size of their output.

We present a algorithm that utilizes both the concept of knowledge transfer, as well as the process behind Generative Adversarial Networks to generate the implicit relations between different modes of data. Our method utilizes Adversarial Network Compression [2] to design the feature map generating backbones for our pedestrian detection networks, which alone can be utilized to support a classification model with multiple modes, however we utilize this model to train *Faster R-CNN* [6] layers to show that multi-modal transfer learning can improve the accuracy of any model.

Thus, our research adds a knowledge-transfer method that utilizes multiple modes that can be applied to increase the accuracy of any deep-network and can be utilized assist in the creation of a highly-accurate, real-time Pedestrian Detection Network.

## 2. Related Work

Both neural network compression and multi-modal learning have been used before in x and y, but synthesizing both to improve small models has not yet been applied in the literature. Our work seeks to combine these methods in order to produce faster, more accurate models that are less computationally expensive.

### 2.1. Adversarial Network Compression

Adversarial Network Compression, as described by *Belagiannis, et al.* is an application of adversarial learning that trains networks by comparing the feature maps of a pretrained teacher network to an untrained student network and updating the weights of the student network such that it produces very similar feature maps to the teacher. We have applied a very similar adversarial model, but made alterations to support multi-modal teachers and separately train *Faster R-CNN* layers following adversarially training our model. We also apply applied the concept of pruning to create smaller models that are able to obtain better results than networks trained with labels.

### 2.2. Multiple Modalities

As seen in *König et al.*, combining visual-optical and infrared information is more effective for classifying pedestrian than visual images alone. Much like the previous work, we utilize the *KAIST Multispectral Pedestrian Benchmark*[4] in order to obtain both the visual and thermal data to both train and test our models. For this research we changed the training data from sets 1-5 to sets 1-6 and the validation/testing data from sets 5-9 to sets 6-9 in order to increase the amount of training information. We also utilize all annotations in the training, validation, and testing of our models.


Figure 1. Qualitative results for four different models are shown for the same frame of the KAIST Dataset. The Fusion model used for the Visual/Infrared image was the ResNet-50 EarlyFusion Model.

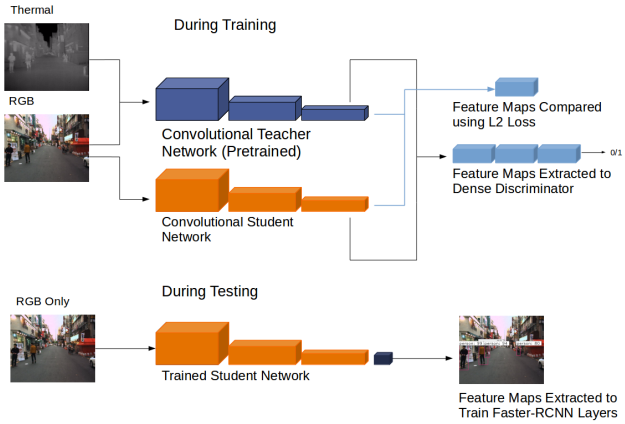


Figure 2. Our method, much like adversarial network compression consists of a teacher and student network. However, instead of two networks being solely of different sizes, on of the networks has more inputs (modes) than the other, which assists in improving accuracy.

### 2.3. Faster R-CNN

For this research we utilized the Faster R-CNN deep network in order to train and test our student models on the KAIST dataset. Although previous work shows that removing the classifier can greatly improve the speed of the classification for Faster R-CNN[8], we decided against using this in our work as it would decrease the accuracy of our classifications. Our Faster R-CNN layers are the full Faster R-CNN model trained for binary detection and classification.

## 3. Method

We propose utilizing adversarial network compression to transfer information about thermal data that are not in the student model's training data in order to improve its accuracy in visual data. In this section, we describe the exact implementation we utilized in order to allow for replication of our experiment.

### 3.1. Models

For our method we decided to focus on two "backbone" models for the Faster R-CNN pedestrian detection: *ResNet-50* [**?**] and *VGG-16* [7]. *ResNet-50* was chosen as multiple-modalities was applied to it in *König et al.* by adding modifications to the first few layers of the model, which gave us a state-of-the-art method for combining the visual and thermal information. VGG-16 was chosen because it is noticeably smaller than ResNet-50 can be adversarially trained by average-pooling the ResNet-50 features by .5 when comparing them to the ResNet-50 teacher model.

For this research we created six models to test pedestrian detection on. The first models we trained were our control models: ResNet-50 trained on visual information only and ResNet-50 trained on thermal information only. We theorized that following the results in the prior work that these two models would have mean-average-precision and miss-rates between the lower-bound and upper-bound for pedestrian detection. For our upper-bound we utilized the ResNet-50 model in *König et al.* that has two convolutional layers before an element wise addition as the input to the second-layer of ResNet-50. Our lower-bound was chosen to be the VGG-16 model trained traditionally with visual data. Finally we had two test models of ResNet-50 and VGG-16, trained only using adversarial transfer learning, to be compared with their single modal models in order to numerically show the improvement from transfer learning.

### 3.2. Software

All of our work for this research was written in the Python distribution Keras [3] using a Tensorflow [1] backend and Tensorboard visualizations.

### 3.3. Faster R-CNN Setup

For training Faster R-CNN using labels on the KAIST dataset, we decided to train each model for 100 groupings of 1000 images at a Learning Rate of .001 using Stochastic Gradient Descent as our Optimizer. Each model was trained using annotations in Pascal VOC format. We also utilized horizontal flipping on half of the images fed into the model to decrease the amount of overfitting. We edited the data such that it would only have once class, "person" for binary classification (KAIST normally has four classes for different people seen in the images). Finally we kept the same

| Metric | ResNet-50 (V) | VGG-16 (V) | ResNet-50 (T) | ResNet-50 EarlyFusion (V/T) | ResNet-50 MidFusion (V/T) |
|---|---|---|---|---|---|
| mAP | 54.75% | 56.76% | 55.24% | **57.37%** | 50.27% |
| Miss Rate | 56.40% | 57.14% | 42.47% | 49.04% | **39.90%** |

Table 1. Quantitative Results

number of images for validation as we did training and only saved our weights when the summed losses of every layer of the Faster R-CNN implementation decreased.

### 3.4. Multi-Modal Transfer Learning

Our model is seen on the next pages in Figure 2. Our model applies the same concepts as the models in *Belagiannis, et al.* for adversarial knowledge transfer. A discriminator that consists of three fully-connected layers (128-256-128) and a sigmoid output is used to compare the teacher and student feature maps, with a label of one signifying the teacher and a layer of zero signifying the student. We then flip the labels in order to trick the discriminator and delay the end of the adversarial game, in this case labeling the student feature map with a label of one and the student feature maps with a dropout of .5 with a label of zero. Finally we feed the teacher feature map into he student model as the "ground truth" and set the student model to decrease the L2 loss between its output and the teacher. Weights are saved only when the L2 loss between the student and teacher decreases over an epoch.

### 3.5. Learning and Optimization

For our adversarial training, we found that the best results for the Adversarial Network occurs when we add Kernel Regularizers with weight decay .002 to all layers with trainable parameters in both the Student, Teacher, and Discriminator Models. This ensures that the game between the Discriminator and Student model can be played long enough for the Student Model to converge to the output of the Teacher Model. For our Adversarial Transfer Model, we set the optimizer to be Adam set to .001 for training the output of the Student Model based off of the Teacher Model and used Stochastic Gradient Descent with a learning rate of .001 and a momentum of .9 for the Discriminator.

For training the Faster R-CNN model on the output of the student feature map, we found that it was best to learning scheme from training the Teacher Model. As such, we also trained it on 100 Epochs of 1000 Images with horizontal flipping on half of the images. As before, we utilized the modified KAIST training data, adding data set 6 to the training data.

## 4. Experiments and Results

We were able to run the experiments on the two control models, the lower-bound, and the upper bound during the Summer. The results are seen in Table 1, as stated before these were trained purely on the KAIST dataset after 100 "Epochs" of 1000 images. As seen above, the most "accurate" model as measured by mean Average Precision was the EarlyFusion ResNet Model, while the model with the lowest miss rate was the MidFusion Model. Our results then do show that using both visual and thermal data as modes is more advantageous than just using one or the other. However it should be noted that the MidFusion model had the lowest mean Average Precision, do to having many false positives, so for the purposes of the Adversarial Training we will be utilizing the EarlyFusion model as our Teacher Model.

The adversarial script for this research was written, however an Adversarial Student Model was unable to be made during the time-frame of the REU and has instead moved to the Future-Work to complete after the REU. Also, looking at the results for the traditionally trained models, it appears that the ResNet-50 models did not converge during training, while the VGG-16 Model did, this is to be expected with Pedestrian Detection and will also have to be addressed in the future work.

## 5. Future Work and Applications

Following the research completed in the REU, we will be looking into creating a student ResNet-50 Model using the adversarial script. If this works as intended, we will utilize the adversarial script with the same Teacher model, but using VGG-16 as the student in order to show that we can improve the results of the smaller, faster model. If these results are successful, the model will further be compressed for more accurate, real-time detection using a MobileNet and tested with our own real-time visual-only pedestrian detection.

## References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] V. Belagiannis, A. Farshad, and F. Galasso. Adversarial network compression. *CoRR*, abs/1803.10750, 2018.

[3] F. Chollet et al. Keras. `https://keras.io`, 2015.

[4] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon. Multispectral pedestrian detection: Benchmark dataset and baselines. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] D. König, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch. Fully convolutional region proposal networks for multispectral person detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 243–250, July 2017.

[6] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[8] L. Zhang, L. Lin, X. Liang, and K. He. Is faster r-cnn doing well for pedestrian detection? *arXiv preprint arXiv:1607.07032*, 2016.