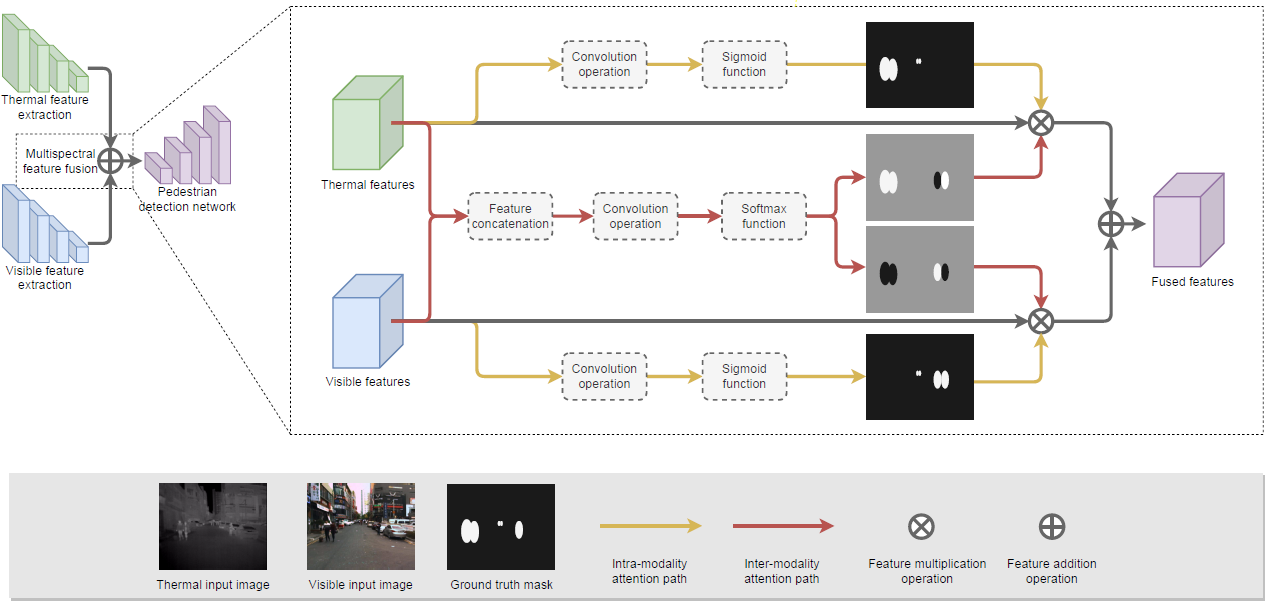
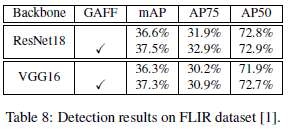
**RGBT Fusion:**

21-01-Guided attentive feature fusion for multispectral **pedestrian detection**

* Intra- and intermodality attention modules help the network to learn adaptive weighing and fusion of multispectral features (i.e. it ideally means that the network give more weight to RGB when it can have more performance and to IR when it is night for ex.)
* The first paper that regards the multispectral feature fusion as a sub-task in the network optimization
* Compared to common feature fusion methods (addition or concatenation), GAFF brings important accuracy gains at low computational cost.
* Both [11] and [6] use the illumination information as a clue for the adaptive fusion: they train a separate network to estimate the illumination value from a given image pair, then [11] uses the predicted illumination value to weigh the detection results from both the thermal and visible images. [6] uses the illumination value to weigh the detection results from a day illumination sub-network and a night illumination sub-network.
* ↑ such handcraft weighting is limited and produce sub-optimal performance

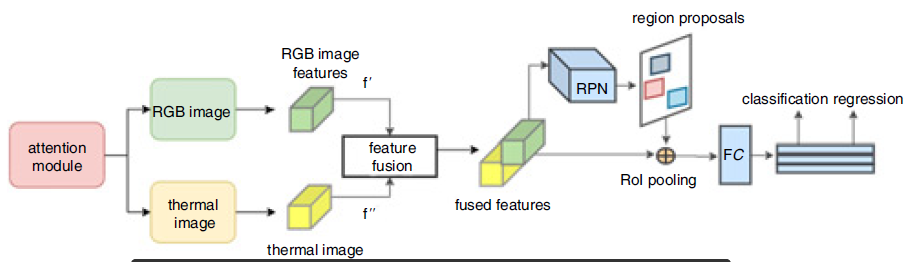


* We can also have 3 path. One for IR only features, one for RGB only and one constantly mixing as it goes one (3 encoder, 3 detector → Blindness detection)
* Only pedastrians taller than 50 Pixel are considered → Evaluation code from: https://github.com/Li- Chengyang/MSDS-RCNN/tree/master/lib/datasets/KAISTdevkit-matlabwrapper
* KAIST → 9.8k well-alighned image pairs for pedestrians. [10] proposes a ”sanitized” version of the annotations, where numerous annotation errors are removed. (Any Dataset could be labeled for IR by Yolo)
* FLIR → 5k pair images for 3 category (car, person and bicycle)
* RetinaNet is choosen as base detector
* ResNet18 or VGG16 pre-trained on ImageNet is adopted as the backbone network
* Input image resolution is fixed to 640\*512 for training and evaluation
* Focal loss and balanced L1 loss are adopted as the classification loss and the bb regression loss
* 72.9% AP50 on FLIR dataset

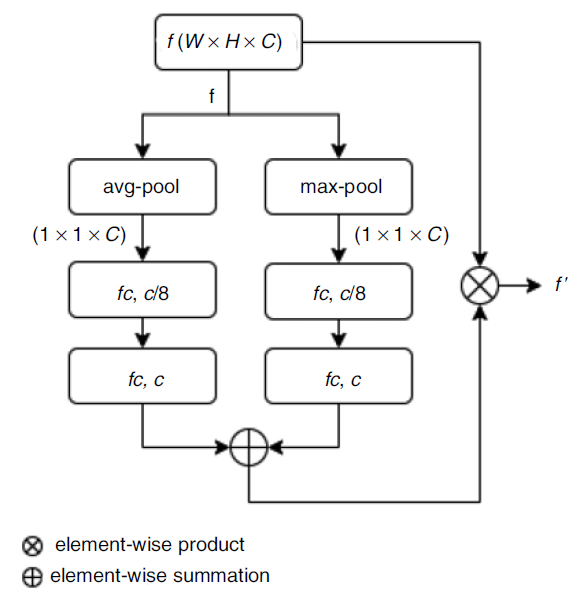


20-11-Attention-based bi-modal fusion:

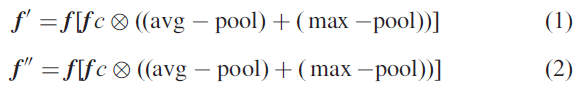
* Improved mAP on FLIR dataset by 4.13 points
* Faster RCNN as detector
* ResNet 101 as feature extractor
* Mid-Fusion (feature level)
* Have a look on DenseFuse paper, where preserved features help classification
* Fusion method and feature map attention details vary with the architecture, as features look different using different methods and architecture
* The primary idea is to take advantage of RGB networks and learn the features from the ROI



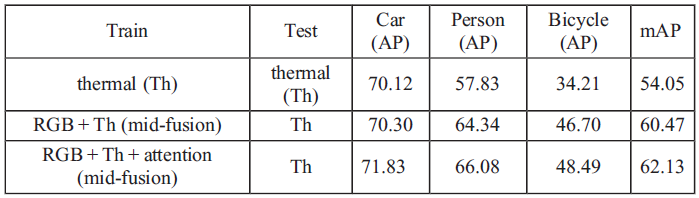
* Attention module is inspired by “Multi-Domain Attentive Detection Network” paper. It prioritize data from different sensors by giving them weight
* Attention module structure is shown below:



* Attention module is placed in every resnet block after the relu and first batch normalization
* Therefore, I do not think that the attention module representation is placed correctly in the first figure!

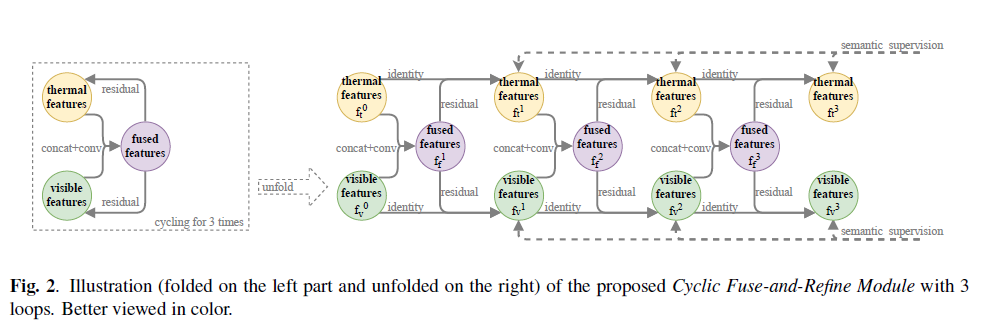


* (1) and (2) then are stacked together in a single feature vector and fed into the RPN of the Faster-RCNN.
* Late fusion is also tried but mid level fusion performs better. (Late fusion means pooled ROI features of thermal and RGB were obtained separately from their respective region proposals and then concatenated together)
* For training details read the paper
* Faster-RCNN is used with pre trained weights on COCO dataset
* FLIR baseline is 58.0 mAP (at 0.5 IoU) and “Borrow from anywhere” provide 61.54 mAP
* The table below shows the results of three training mode

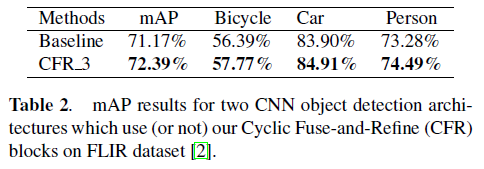


20-09-Multispectral fusion for obj. detection with cyclic fuse and refine blocks:

* The main idea is to refine the mono-spectral features with the fused multispectral features multiple times consecutively in the network

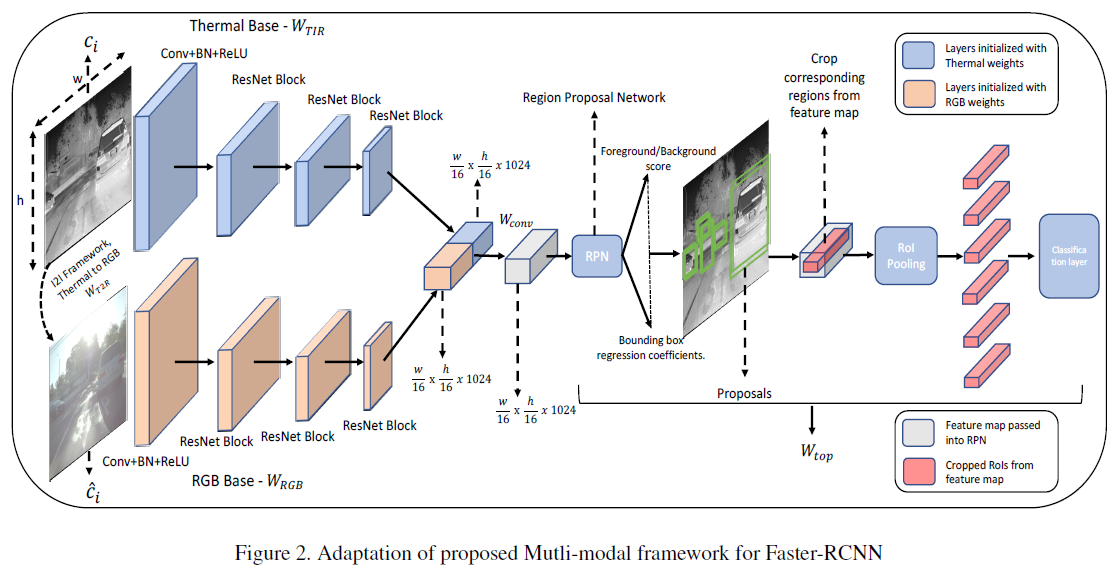


* Results:

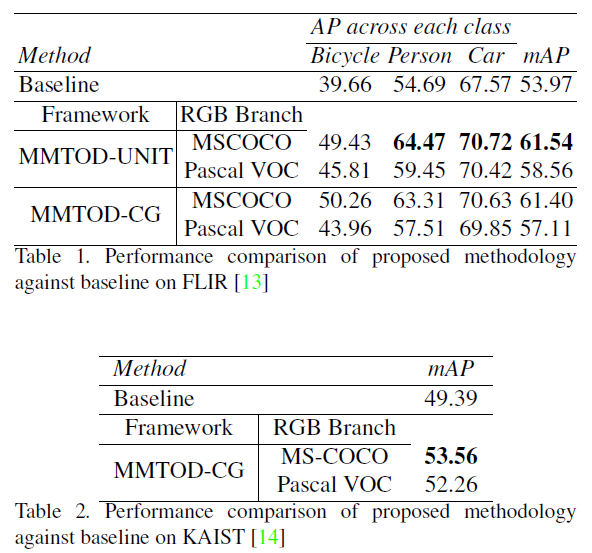


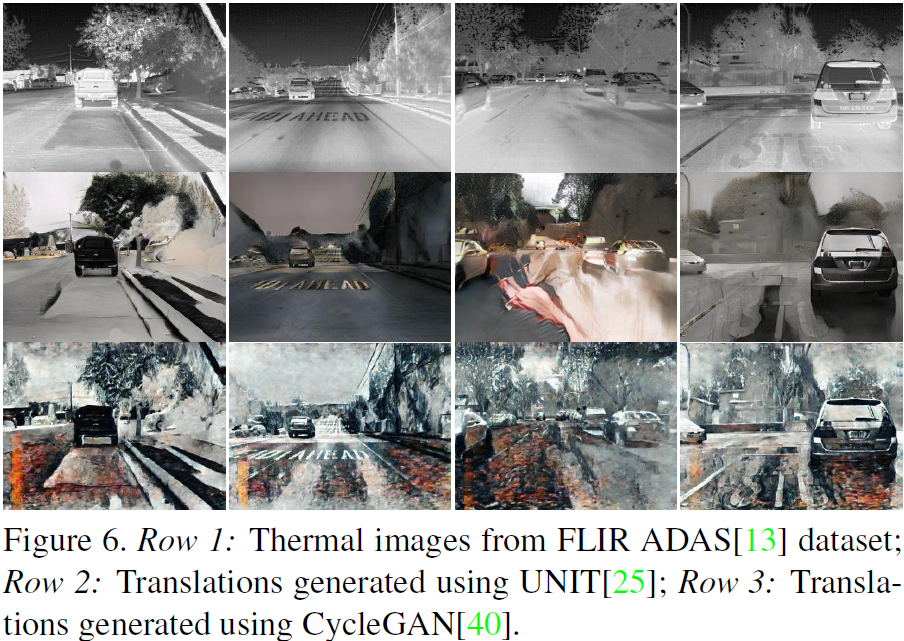
20-07-Borrow from anywhere:

* Object detection networks are trained on large-scale RGB datasets such as ImageNet, PASCAL-VOC and COCO but they are not available in IR domain.
* This paper propose a network with two branches. One branch is pre-trained on above datasets and fine tuned using RGB image which is translated from IR by I2I (image-to-image) framework
* The second branch is for IR



* The key idea is to borrow knowledge from data-rich domains such as visual (RGB) without explicit need for a paired multimodal dataset.
* The objective is achieved by I2I translation methods (UNIT and CycleGAN), which translate any given IR image to RGB image and make training without image pairs possible
* The method could beat FLIR baseline results using only a fourth of the dataset
* Each branch is initialized with a model pre-trained on images from that domain
* FLIR and KAIST is used for IR branch but without RGB images; as the RGB images will be produced by I2I methods
* So to summaries: IR branch is trained end-to-end with IR images. RGB branch is trained on COCO and PASCAL VOC for the same classes. Both branches have weight now fine tuning start from here with IR and fake RGB images

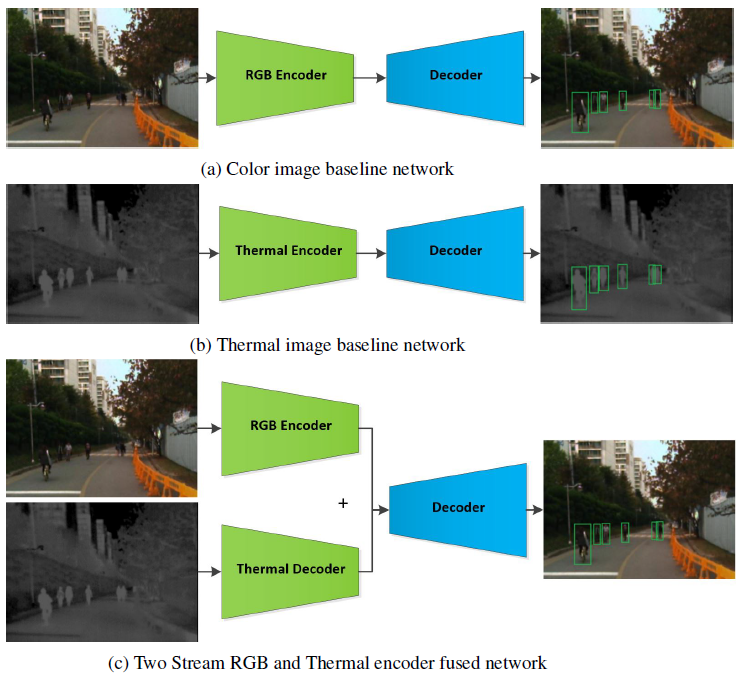




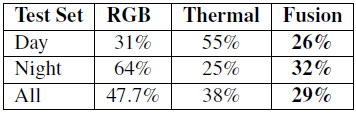
* Github code : <https://github.com/tdchaitanya/MMTOD>

20-07-CNN based color and thermal image fusion:

* (c) is the proposed fusion.
* Faster-RCNN for object detection
* VGG16 as encoder
* Features generated by encoder are fed to RPN (region proposal network)
* ROI are generated from RPN and fed to ROI pooling and classification
* (a) and (b) are trained to be compared with (c)

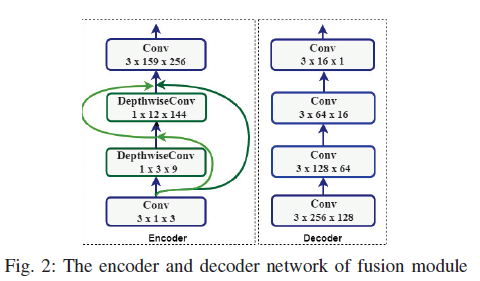


* By a separate thermal decoder, the goal is to be able to combine the existing visual object detection of the car with the additional pipeline and therefore do not touch the RGB encoder
* It also leave the camera network to be free and used for multiple tasks
* Output volume after the summation junction is of the same shape as the feature vector generated by each decoder alone
* The decoder in all a,b and c versions are the same
* Disadv. → 2 parallel networks are naturally more computational expensive than version a, b
* Adv. → RGB Encoder can also be used for other tasks
* KAIST Dataset: Every second image is deleted from the training set as they are redundant. Also object with less than 50 pixels are removed as they are less informative and increase false positives. KAIST has only 1 class (pedastrians)
* FLIR Dataset: Different cameras are used ranging from 0.3 to 3.1 MP with different FOV. Thermal intensity imaging has a bit-depth of 14-bits and 8-bits resolution | FLIR is very challenging to be used for fusion algorithm! They struggled with resolution/aspect ratio problem of FLIR dataset (640\*512 vs. 1800\*1600) | 14-bit input resolution led to fully leverage weights of pre-trained 8-bit network initialization. | sensor extrinsic calibration is an unsolved problem
* Read the paper for detailed info about hyperparameters and training setup
* Log average miss rate is used as the performance metric
* The table below compares all three versions a,b and c in day and night condition. “%” shows the log average miss rate.



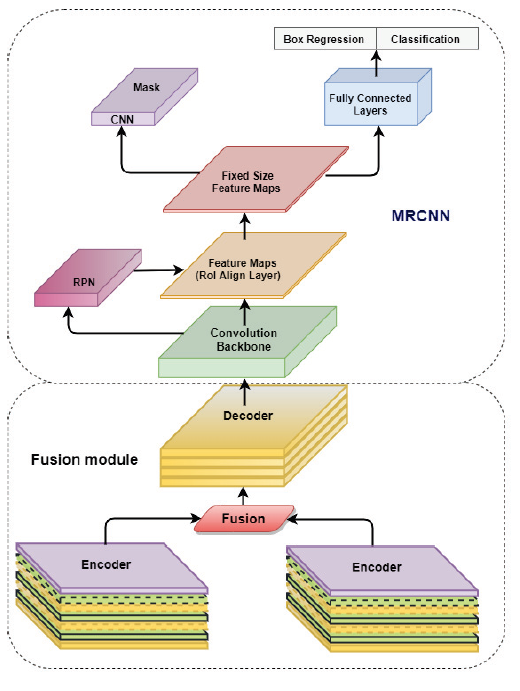
**20-06-Night Vision Surveillance: Obj. Detection using thermal and visible images:**

* Introduce the past fusion techniques in recent years
* Claims that DeepFuse is the best one among them but due to inadequate salient feature extraction, fails to preserve meaningful information
* Therefore, authoes in DenseFuse, introduced dense layers but this layers add computational complexity → not real time efficient
* **This paper propose depthwise convolution utilization in encode to reduce complexity (i.e. depth separable convolution)**
* [**https://medium.com/@zurister/depth-wise-convolution-and-depth-wise-separable-convolution-37346565d4ec**](https://medium.com/@zurister/depth-wise-convolution-and-depth-wise-separable-convolution-37346565d4ec)

****

* It is reviewed that pedestrian detection during the night has higher miss rate without IR
* **1\*1 filter in depthwise Conv. + previous features are concatenated and passed to next features**
* **After feature extraction, weighted sum fusion strategy is adopted to fuse the features based on the illumination factor of each image.** We might do the same for final Detection

****



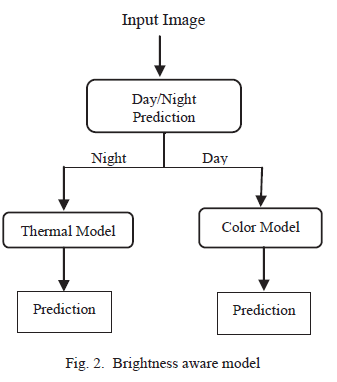
* Decoder reconstruct the fused image
* It is observed that, if the network is trained only on IR images, due to limited learned features, it does not perform well, therefore they have decided to train on COCO first to shape their filters and then train on FLIR for tuning the filters

19-09-(Yolov2)Multi-domain attentive detection network:

* Not much to say

19-04-Deep learning based pedestrian detection at all light conditions:

* Brightness aware network to decide between day → RGN or night → IR



Classification:

* First method: Prediction of Day/Night by MobileNet
* Second method: By Image processing
* mAP of Image processing method turned out to be more than CNN method

Detection:

* RPN (region proposal network) is trained with FLIR dataset for thermal images and PASCAL VOC for color images

17-11-(Yolov2)Deep obj. classification:

* While applying transfer learning to take advantage of low level features, weights need to be adapted to compensate for the different size of images
* Pre-trained network is fine-tuned by IR frames