# Original Objective:

The focus of this research is to develop a transformer model that integrates the information from different modalities together to enhance the prediction as well as address the challenges posed by missing modalities.

# Ideas:

We use the concept like retrieval from database by using query, keys, and value utilizing attention calculation in transformers. We use query (a query we wish to run on a database) from one modality and keys (the keys to search on in the database) and values (values corresponding to each key in the database) from other modalities. Cross-attention in transformer encoder is used to gain context from another modality/ input type as a method of TokenFusion in channels. This is accomplished by pairwise exchange of keys and values from different modalities. For example, to gain context from text for object detection, we simply extract the queries matrix from text modality, and keys and values matrix from the RGB and IR modalities. Moreover, self-attention blocks at the end of our model architecture would allow the model to further process the combined representations as well as enable the model to understand the dependencies between different parts of the input from different modalities.

# Dataset:

95k color-thermal pairs (640x480, 20Hz) captured from a vehicle make up the KAIST Multispectral Pedestrian Dataset. Each pair (person, people, cyclists) has 1,182 distinct pedestrians and 103,128 dense annotations overall, all of which are carefully annotated. Temporal association between bounding boxes, akin to the Caltech Pedestrian Dataset, is included in the annotation.

|  |  |  |  |
| --- | --- | --- | --- |
| Name of the Database | Image  Type | RGB | Thermal |
| The KAIST Multispectral Pedestrian Dataset | # of data | 95,328 | 95,328 |
| Size | 640 x 480 | 640 x 480 |
| # Filtered image with good alignment | 95,328 | 95,328 |
| # further alignment | 4750 | 4750 |
| Objective of the Training: | Pedestrian detection using labeled images (bounding box - class label and occlusion labels). | |

More information can be found in CVPR 2015 paper. **Paper link**: [CVPR15\_Pedestrian\_Benchmark.pdf (soonminhwang.github.io)](https://soonminhwang.github.io/rgbt-ped-detection/misc/CVPR15_Pedestrian_Benchmark.pdf)

**Dataset Link**: <https://soonminhwang.github.io/rgbt-ped-detection/data/>

Three types of annotations are available in the KAIST Multispectral Pedestrian Dataset for training. First, Hwang et al. [1] gave the original annotations. Second, Li et al. [2] supplied the cleaned annotations. Finally, Zhang et al. [3] gave the paired annotations. We employed cleaned test annotations for evaluation. The sanitized annotations remove any annotation mistakes, such as misclassified, mislocalized, and misaligned regions. Since most recent works have employed annotations for evaluation, we likewise adopt them to do a fair comparison.

Dataset References:

[1] - S. Hwang, J. Park, N. Kim, Y. Choi, and I. Kweon, “Multispectral pedestrian detection: Benchmark dataset and baseline,” in Proc. IEEE Conf. Comput. Vision Pattern Recognit., 2015, pp. 1037–1045

[2] - C. Li, D. Song, R. Tong, and M. Tang, “Multispectral pedestrian detection via simultaneous detection and segmentation,” in Proc. Brit. Mach. Vision Conf., 2018, pp. 225.1–225.12.

[3] - L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, “Weakly aligned cross-modal learning for multispectral pedestrian detection,” in Proc. IEEE Int. Conf. Comput. Vision, 2019, pp. 5126–5136.



# Modifications

Unlike DETR, our focus in this experiment is to develop a transformer model that integrates information from different modalities together to enhance the prediction. Our modified DETR transformer encoders extract the features from RGB and IR modality via two parallel ResNet50 backbone and harness the vast potential of heterogeneous data via channel fusion of modalities features (RGB and IR) using two parallel cross-attention encoders unlike the original DETR transformer model. Afterward, the model performs object detection by using a DETR decoder with object queries.

References:

[1]. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. arXiv preprint arXiv:2005.12872.

[2]. Cai, Z., Vasconcelos, N.: Cascade R-CNN: High quality object detection and instance segmentation. PAMI (2019)

[3]. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: AISTATS (2010)

[4]. He, K., Girshick, R., Doll´ar, P.: Rethinking imagenet pre-training. In: ICCV (2019)

[5]. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)

[6]. Hosang, J.H., Benenson, R., Schiele, B.: Learning non-maximum suppression. In: CVPR (2017)

[7]. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: CVPR (2018)