## 6 Appendix

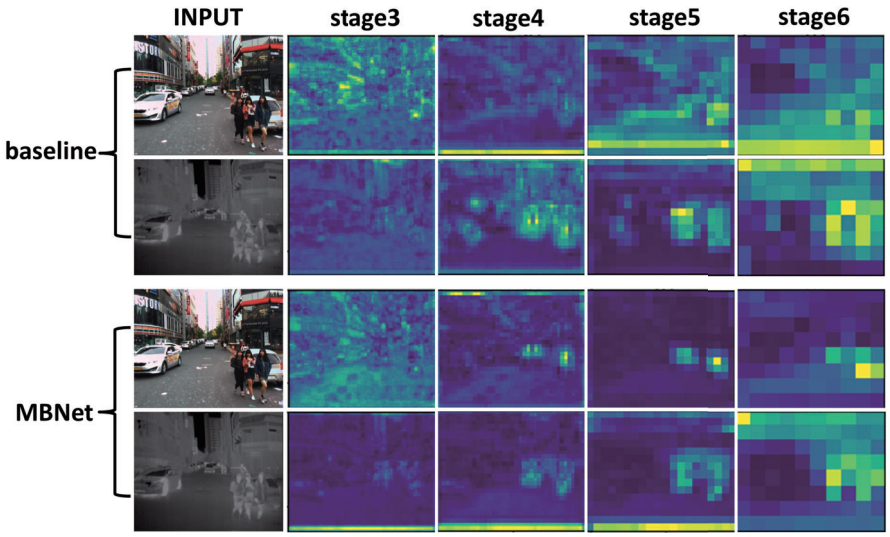


**Fig. 6.** The overall perspective of the baseline and MBNet feature maps ($H \times W \times C$) in stage3, stage4, stage5, stage6.

The DMAF module fuses two modalities at the channel level, so it should be noted that the visualization in Fig. 3 is just one channel ($H \times W \times 1$) in order to have an intuitive understanding of the effect of DMAF. From an overall perspective of the feature maps ($H \times W \times C$) shown in Fig. 6, the pedestrian region features become more salient with DMAF added. In order to have a deeper understanding of the DMAF module, we try to explain the role of the DMAF module from the view of modality redundancy.

we introduce the Pearson product-moment correlation coefficient ($|\rho|$) between two modality feature maps to represent the modality redundancy. Since the feature maps are three-dimensional data ($H \times W \times C$), we divide the feature maps into channel level ($1 \times 1 \times C$) and feature level ($H \times W \times 1$). We randomly select 100 pair-images from the KAIST test set and the Pearson product-moment correlation coefficient ($|\rho|$) is calculated from both levels. The stage3, stage4, stage5, stage6 feature maps in the backbone (which are used to predicts the location and score) from the baseline and MBNet as well as feature maps after modality alignment module are chosen as the experimental samples. We make statistics on the correlation coefficients $|\rho|$ between two modalities and illustrate the proportion of different correlation coefficients $|\rho|$ (0.0$\sim$1.0) as a line chart in Fig. 7.

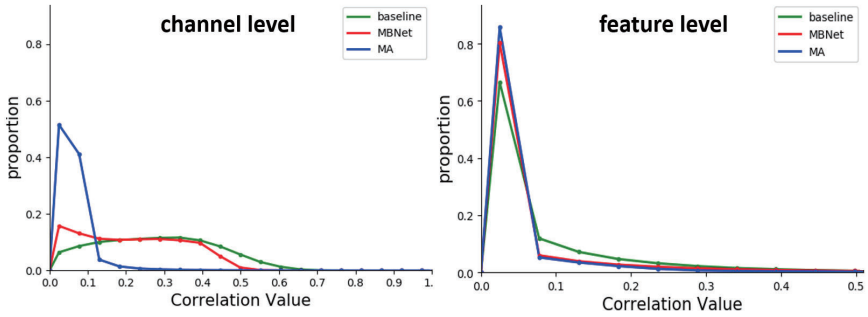


**Fig. 7.** The modality redundancy (Pearson product-moment correlation coefficient $|\rho|$) between two modality feature maps from channel level ($1 \times 1 \times C$) and feature level ($H \times W \times 1$). The green and red line represent the baseline and MBNet stage3–6 feature maps respectively. The blue line represents the feature maps after Modality Alignment (MA) module in MBNet.

With the DMAF module added, two modalities tend to be unrelated from both the channel and feature level (red line vs. green line), which means redundant information is reduced. After the modality alignment module, the correlation between the two modalities is further reduced, especially at the channel level. In our opinion, **the DMAF module facilitates modality interaction in the network which reduces the learning of redundancy and conveys more information**. The effective extraction of useful information and the elimination of redundancy between two modalities are problems worthy of studying in the future.