Supplementary Material for A^{2}RNet: Adversarial Attack Resilient Network for Robust Infrared and Visible Image Fusion

Anonymous submission

Abstract

Here, we provide supplementary materials for the main text. First, we detail the specific process of pseudo-label generation. Then, additional experimental details and results are provided to facilitate a better understanding. All in all, thank you for taking the time to read this document.

Pseudo-label Generation

As shown in Fig. 1, we use a common CNN to generate pseudo-labels. l_1, SSIM, and gradient loss are used as the loss functions during training. We utilize these pseudo-labels as the necessary references for adversarial attacks and training, which introduces a new paradigm for research on adversarial robustness in the IVIF task. Moreover, these pseudo-labels are more "moderate" and less likely to lead to overfitting.

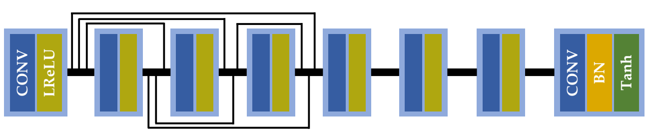

Figure 1: Pipeline of the pseudo-label generation.

More Details of Experiments

Quantitative Metrics

In the quantitative comparison, we choose Entropy (EN), Standard Deviation (SD), Peak Signal-toNoise Ratio (PSNR), Correlation Coefficient (CC) (Shah, Merchant, and Desai 2013) and the Sum of the Correlations of Differences (SCD) (Aslantas and Bendes 2015). Specifically, EN represents the measure of information contained in fused images. SD quantifies the dispersion of pixel values around the mean in fusion results. PSNR reflects the level of distortion by comparing the peak signal power to the noise power. CC represents the degree of linear correlation between source and fused images. SCD denotes the characterization of differences between source and fused images. Clearly, EN and SD are metrics calculated based on the fused results themselves, while PSNR, CC and SCD evaluate the relationship between source images and fused images. For all metrics, higher values indicate better image quality.

Configurations of Downstream Tasks

We select the M^{3}FD (Liu et al. 2022) and MFNet (Ha et al. 2017) datasets, which provide annotated detection and segmentation labels. Note that we use the fusion results as inputs to retrain the downstream task models. For the detection task, we employ YOLOv5 (Redmon et al. 2016) as the detector. The optimizer, learning rate, epochs, and batch size are set to SGD optimizer, 1e-2, 300, and 8, respectively. 4200 image pairs from the M^{3}FD dataset are divided into training, validation, and test sets in an 8:1:1 ratio. In the segmentation task, DeepLabV3+ (Chen et al. 2018) is introduced with 300 epochs and a batch size of 8, keeping the other parameters consistent with the original model. A total of 1,083 fusion results from the MFNet dataset are used as the training set, while the remaining 361 are used as the test set.

Fusion Results under Different Perturbations

For a more comprehensive comparison, we have included an additional set of experiments with 8/255 attacks in the supplementary materials. Note that adversarial examples during the adversarial training are produced using the same perturbations. As shown in Fig. 2, two sets from the M^{3}FD and MFNet datasets are presented. Clearly, our method does not exhibit undesirable artifacts when subjected to stronger perturbations. However, other methods exhibited more pronounced noise or distortions. It can demonstrate that our method maintains robustness even under stronger perturbations.

Quantitative Comparisons in Downsteam tasks

In this supplementary material, we provide detailed quantitative results for each category in the detection and segmentation tasks, where the subscript indicates the degree of change under adversarial conditions. As shown in Table. 1 and 2, our method achieves the best performance across all categories, demonstrating that fused images generated by A^{2}RNet remain robust in downstream task performance. Combined with the qualitative results in the main text, the downstream task performance of our method remains a leading position under attacks.

Figure 2: Fusion comparisons with SOTA methods in MFNet and M[3]FD datasets. We apply PGD to clean samples and add stronger perturbations with $\epsilon = 8/255$ to generate AEs.

Category	AP@.5							
	TarDAL	SeAFusion	IGNet	PAIF	CoCoNet	LRRNet	EMMA	Ours
People	$0.530_{\downarrow 0.171}$	$0.704_{\downarrow 0.041}$	$0.559_{\downarrow 0.134}$	$0.703_{\downarrow 0.034}$	$0.209_{\downarrow 0.133}$	$0.680_{\downarrow 0.038}$	$0.676_{\downarrow 0.032}$	$0.764_{\downarrow 0.011}$
Car	$0.597_{\downarrow 0.187}$	$0.860_{\downarrow 0.040}$	$0.825_{\downarrow 0.042}$	$0.859_{\downarrow 0.029}$	$0.680_{\downarrow 0.063}$	$0.841_{\downarrow 0.024}$	$0.801_{\downarrow 0.039}$	$0.893_{\downarrow 0.001}$
Bus	$0.491_{\downarrow 0.284}$	$0.776_{\downarrow 0.073}$	$0.582_{\downarrow 0.200}$	$0.790_{\downarrow 0.049}$	$0.456_{\downarrow 0.000}$	$0.779_{\downarrow 0.045}$	$0.751_{\downarrow 0.052}$	$0.866_{\downarrow 0.012}$
Motorcycle	$0.348_{\downarrow 0.093}$	$0.584_{\downarrow 0.018}$	$0.411_{\downarrow 0.158}$	$0.554_{\downarrow 0.015}$	$0.203_{\downarrow 0.025}$	$0.620_{\uparrow 0.014}$	$0.600_{\downarrow 0.007}$	$0.647_{\downarrow 0.021}$
Truck	$0.518_{\downarrow 0.123}$	$0.771_{\downarrow 0.050}$	$0.529_{\downarrow 0.205}$	$0.751_{\downarrow 0.080}$	$0.238_{\downarrow 0.200}$	$0.785_{\downarrow 0.028}$	$0.711_{\downarrow 0.054}$	$0.809_{\downarrow 0.021}$
Lamp	$0.289_{\downarrow 0.106}$	$0.699_{\downarrow 0.027}$	$0.436_{\downarrow 0.186}$	$0.631_{\downarrow 0.069}$	$0.077_{\downarrow 0.138}$	$0.518_{\downarrow 0.073}$	$0.635_{\downarrow 0.046}$	$0.712_{\downarrow 0.038}$
mAP@.5	$0.462_{\downarrow 0.161}$	$0.732_{\downarrow 0.042}$	$0.557_{\downarrow 0.154}$	$0.715_{\downarrow 0.046}$	$0.311_{\downarrow 0.093}$	$0.704_{\downarrow 0.032}$	$0.696_{\downarrow 0.038}$	$0.781_{\downarrow 0.024}$

Table 1: Quantitative results of detection. Red and blue denote the optimal and suboptimal results, respectively. The subscripts indicate the change compared adversarial conditions with the clean.

Category	IoU							
	TarDAL	SeAFusion	IGNet	PAIF	CoCoNet	LRRNet	EMMA	Ours
Background	$0.963_{\downarrow 0.005}$	$0.971_{\downarrow 0.004}$	$0.967_{\downarrow 0.014}$	$0.974_{\downarrow 0.004}$	$0.962_{\downarrow 0.014}$	$0.978_{\downarrow 0.000}$	$0.973_{\downarrow 0.010}$	$0.979_{\downarrow 0.005}$
Car	$0.685_{\downarrow 0.058}$	$0.802_{\downarrow 0.024}$	$0.749_{\downarrow 0.127}$	$0.831_{\downarrow 0.056}$	$0.704_{\downarrow 0.041}$	$0.792_{\downarrow 0.057}$	$0.808_{\downarrow 0.086}$	$0.845_{\downarrow 0.055}$
Person	$0.663_{\downarrow 0.040}$	$0.634_{\downarrow 0.064}$	$0.469_{\downarrow 0.213}$	$0.643_{\downarrow 0.078}$	$0.383_{\downarrow 0.113}$	$0.615_{\downarrow 0.099}$	$0.625_{\downarrow 0.106}$	$0.686_{\downarrow 0.040}$
Bike	$0.340_{\downarrow 0.011}$	$0.570_{\downarrow 0.083}$	$0.511_{\downarrow 0.169}$	$0.547_{\downarrow 0.132}$	$0.434_{\downarrow 0.096}$	$0.597_{\downarrow 0.009}$	$0.591_{\downarrow 0.094}$	$0.645_{\downarrow 0.053}$
Curve	$0.136_{\downarrow 0.145}$	$0.449_{\downarrow 0.089}$	$0.281_{\downarrow 0.249}$	$0.306_{\downarrow 0.201}$	$0.166_{\downarrow 0.065}$	$0.427_{\downarrow 0.012}$	$0.488_{\downarrow 0.092}$	$0.508_{\downarrow 0.091}$
Car Stop	$0.252_{\downarrow 0.094}$	$0.600_{\downarrow 0.042}$	$0.462_{\downarrow 0.198}$	$0.492_{\downarrow 0.157}$	$0.294_{\downarrow 0.092}$	$0.594_{\downarrow 0.009}$	$0.676_{\downarrow 0.021}$	$0.685_{\downarrow 0.031}$
Guardrail	$0.153_{\downarrow 0.656}$	$0.626_{\downarrow 0.066}$	$0.594_{\downarrow 0.137}$	$0.602_{\downarrow 0.129}$	$0.295_{\downarrow 0.049}$	$0.663_{\downarrow 0.015}$	$0.645_{\downarrow 0.119}$	$0.682_{\downarrow 0.087}$
Color Cone	$0.427_{\downarrow 0.084}$	$0.509_{\downarrow 0.053}$	$0.418_{\downarrow 0.210}$	$0.506_{\downarrow 0.100}$	$0.257_{\downarrow 0.125}$	$0.520_{\downarrow 0.007}$	$0.581_{\downarrow 0.034}$	$0.604_{\downarrow 0.041}$
Bump	$0.256_{\downarrow 0.116}$	$0.405_{\downarrow 0.270}$	$0.176_{\downarrow 0.335}$	$0.302_{\downarrow 0.342}$	$0.192_{\downarrow 0.135}$	$0.422_{\downarrow 0.044}$	$0.453_{\downarrow 0.210}$	$0.461_{\downarrow 0.205}$
mIoU	$0.415_{\downarrow 0.069}$	$0.618_{\downarrow 0.078}$	$0.514_{\downarrow 0.184}$	$0.578_{\downarrow 0.134}$	$0.409_{\downarrow 0.080}$	$0.623_{\downarrow 0.027}$	$0.649_{\downarrow 0.086}$	$0.677_{\downarrow 0.068}$

Table 2: Quantitative results of segmentation. Red and blue denote the optimal and suboptimal results, respectively. The subscripts indicate the change compared adversarial conditions with the clean.

References

Aslantas, V.; and Bendes, E. 2015. A new image quality metric for image fusion: The sum of the correlations of differences. *Aeu-international Journal of electronics and communications*, 69(12): 1890–1896.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Ha, Q.; Watanabe, K.; Karasawa, T.; Ushiku, Y.; and Harada, T. 2017. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5108–5115. IEEE.

Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; and Luo, Z. 2022. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5802–5811.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Shah, P.; Merchant, S. N.; and Desai, U. B. 2013. Multifocus and multispectral image fusion based on pixel significance using multiresolution decomposition. *Signal, Image and Video Processing*, 7: 95–109.