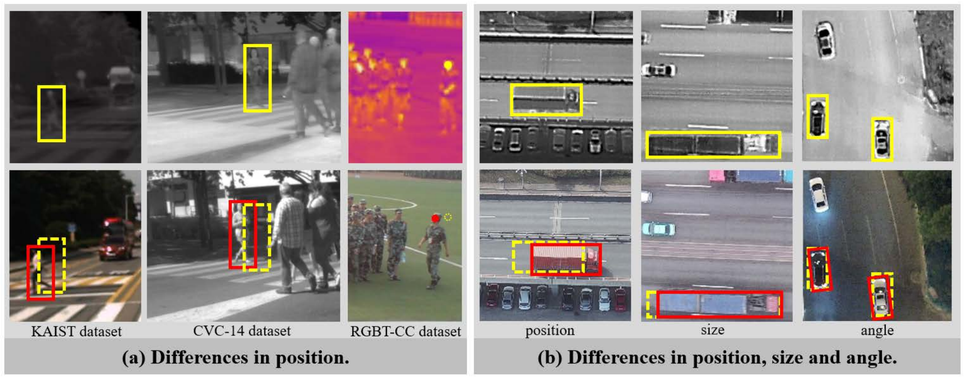
Translation, Scale and Rotation: Cross-Modal Alignment Meets RGB-Infrared Vehicle Detection

Maoxun Yuan, Yinyan Wang, Xingxing Wei

Beihang University

Problem Presentation and Contribution

The same objects on image pairs differs in location, scale and angle.

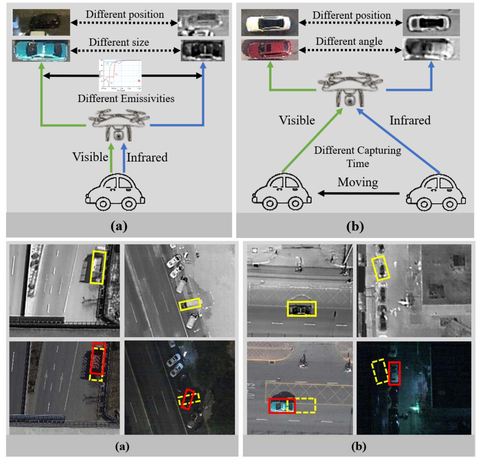


(a) Differences in position. (b) Differences in position, size and angle.

Our contributions:

- The cross-modal weakly misalignment problem specific to the RGB-IR object detection in aerial images is presented.
- A Translation-Scale-Rotation Alignment (TSRA) module is proposed to align the feature maps of two modality objects.

Cross-modal Misalignment Problem Analysis



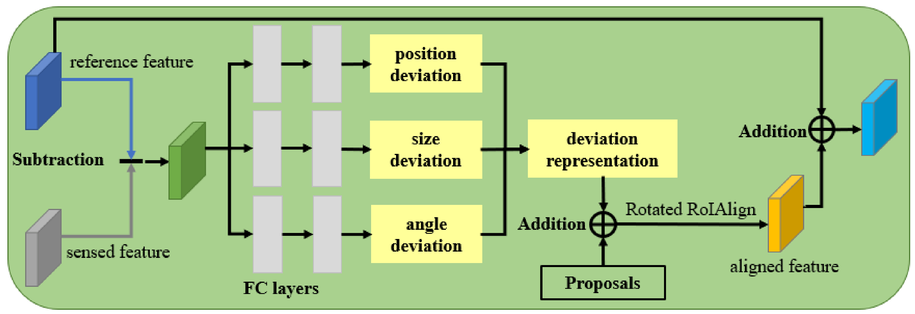
Hardware Errors: Hardware errors are mainly reflected in radiation distortions (first row (a)) and clock skews (first row (b)). The radiation differences will cause images have different representations for the same objects. The clock skew between two sensors can lead to pixel-misalignment of image pairs, especially for locally moving objects such as cars on a highway .

Annotation Errors: In the process of multispectral data annotation, it is difficult to ensure that the objects are annotated accurately in different modalities (second row (a)). Hardware errors and annotation errors can occur simultaneously in the same object (second row (b)).

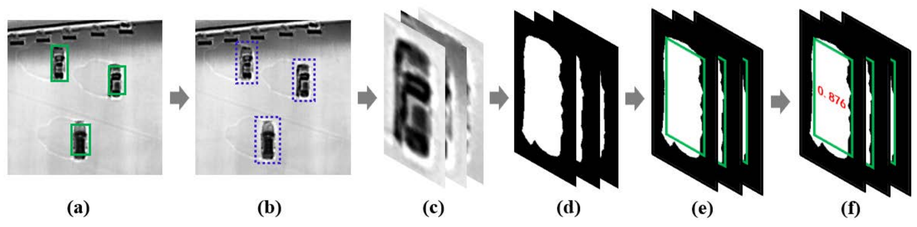
TSRA Module

TSRA module mainly consists of two parts: Alignment Process and Modality-Selection Strategy.

Alignment Process

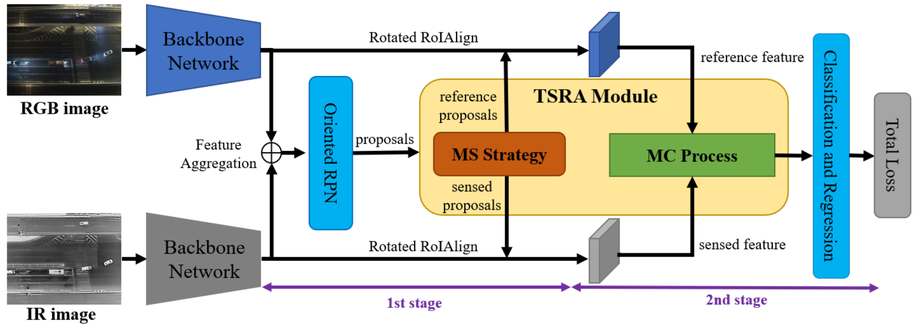


Modality-Selection Strategy



TSRA-based Object Detectors

Overall Architecture



Loss Function

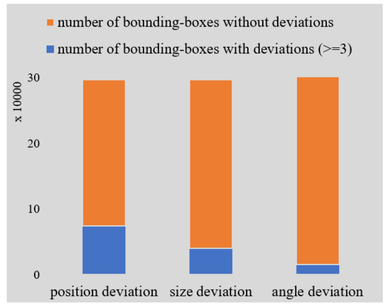
$$L_{deviation}(\{p_i^*\},\{t_i\},\{t_i^*\},\{s_i\},\{s_i^*\},\{r_i\},\{r_i^*\}) = \frac{1}{N_{deviation}}\sum_{i=1}^{n} p_i^*(smoothL_1(t_i - t_i^*) + smoothL_1(s_i - s_i^*) + smoothL_1(r_i - r_i^*))$$

$$t_x^* = ((x_s - x_r)\cos\theta_r + (y_s - y_r)\sin\theta_r)/w_r,$$
$$t_y^* = ((y_s - y_r)\cos\theta_r + (x_s - x_r)\sin\theta_r)/h_r,$$
$$s_w^* = \log w_s / w_r, s_h^* = \log h_s / h_r,$$
$$r_\theta^* = ((\theta_s - \theta_r) \bmod 2\pi)/2\pi$$

$$L = L_{cls} + L_{reg} + L_{rpn} + \lambda L_{deviation}$$

Experimental Results

Statistics



As illustrated in above, more than 20% of the bounding-boxes have the deviation problem in our used public DroneVehicle.

Ablation Studies

Table 1. Ablation experiments of TSRA module on DroneVehicle dataset. The symbols of 'P', 'S', and 'A' represent Position, Size and Angle prediction branches respectively.

	P	S	A	car	freight-car	truck	bus	van	mAP
TSFADet (full)	✓	✓	✓	**89.88**	**67.87**	**63.74**	**89.81**	**53.99**	**73.06**
TSFADet w/o MJ	✓	✓	✓	89.34	66.53	62.65	89.62	53.67	72.36
TSFADet w/o MS	✓	✓	✓	89.78	66.26	61.44	89.60	53.17	72.05
TSFADet w/o MJ and MS	✓	✓		89.69	65.11	60.39	89.43	51.01	71.13
	✓		✓	89.65	64.77	61.28	89.26	48.57	70.71
		✓	✓	89.68	62.97	60.22	88.90	49.59	70.27
			✓	89.56	62.83	58.35	89.46	47.26	69.49
Baseline				89.45	62.14	57.00	89.09	45.43	68.62

Table 2. Quantitative comparisons of using different methods to demonstrate the contribution of the MS strategy.

Methods	car	truck	freight-car	bus	van	mAP
RGB Modality	89.47	65.56	60.36	89.63	52.82	71.57
IR Modality	89.78	66.26	61.44	89.60	53.17	72.05
Random strategy	89.53	66.71	62.38	89.74	53.74	72.42
MS strategy	**89.88**	**67.87**	**63.74**	**89.81**	**53.99**	**73.06**

Comparisons

Table 3. Evaluation results on the DroneVehicle dataset. The last column refers to input modalities of the approach.

Detectors	car	truck	freight-car	bus	van	mAP	Modality
Faster R-CNN(OBB) [22]	79.69	41.99	33.99	76.94	37.68	54.06	RGB
RetinaNet(OBB) [14]	78.45	34.39	24.14	69.75	28.82	47.11	
ROI Transformer [5]	61.55	55.05	42.26	85.48	44.84	61.55	
S^2ANet [7]	79.86	50.02	36.21	82.77	37.52	57.28	
Oriented R-CNN [28]	80.26	55.39	42.12	86.84	46.92	62.30	
Faster-R-CNN(OBB) [22]	89.68	40.95	43.10	86.32	41.21	60.27	IR
RetinaNet(OBB) [14]	88.81	35.43	39.47	76.45	32.12	54.45	
ROI Transformer [5]	89.64	50.98	53.42	88.86	44.47	65.47	
S^2ANet [7]	89.71	51.03	50.27	88.97	44.03	64.80	
Oriented R-CNN [28]	89.63	53.92	53.86	89.15	40.95	65.50	
Halfway Fusion(OBB) [15]	89.85	60.34	55.51	88.97	46.28	68.19	RGB +IR
CIAN(OBB) [36]	89.98	62.47	60.22	88.9	49.59	70.23	
AR-CNN(OBB) [37]	**90.08**	64.82	62.12	89.38	51.51	71.58	
TSFADet(Ours)	89.88	67.87	63.74	**89.81**	53.99	73.06	
Cascade-TSFADet (Ours)	90.01	**69.15**	**65.45**	89.70	**55.19**	**73.90**	

Table 4. Speed versus accuracy on the DroneVehicle dataset.

Method	FPS	mAP	Input	framework
Halfway Fusion(OBB)	20.4	68.19	RGB+IR	two-stage
CIAN(OBB)	**21.7**	70.23	RGB+IR	one-stage
AR-CNN(OBB)	18.2	71.58	RGB+IR	two-stage
TSFADet(Ours)	18.6	**73.06**	RGB+IR	two-stage