

Enhancing Mid–Low-Resolution Ship Detection With High-Resolution Feature Distillation

Shitian He[✉], Huanxin Zou[✉], Yingqian Wang[✉], Runlin Li, Fei Cheng, Xu Cao, and Meilin Li

Abstract—To enhance mid–low-resolution ship detection, existing methods generally use image super-resolution (SR) as a preprocessing step and feed the super-resolved images to the detectors. However, these methods only use high-resolution (HR) images as ground-truth labels to supervise the training of their SR module but overlook the rich HR information in the detection stage. Inspired by the recent advances in knowledge distillation, in this letter, we design a feature distillation framework to fully exploit the information in ground-truth HR images to handle mid–low-resolution ship detection. Our framework consists of a student network and a teacher network. The student network first super-resolves input images using an SR module and then feeds the super-resolved images to the detection module. The teacher network whose architecture is the same as the student detection module directly takes HR images as input to generate HR feature representation and then distills these HR features to the student network through a distillation loss. Using our feature distillation framework, HR images are not only used as ground-truth labels to train the SR module but also provide “ground-truth” features to train the detection module, which enhances the detection performance of the student network. We apply our framework to several popular detectors, including *FCOS*, *Faster-RCNN*, *Mask-RCNN*, and *Cascade-RCNN*, and conduct extensive ablation studies to validate its effectiveness and generality. Experimental results on the HRSC2016, DOTA, and NWPU VHR-10 datasets demonstrate that, when applying our framework to *Faster-RCNN*, our method can outperform several state-of-the-art detection methods in terms of mAP50 and mAP75.

Index Terms—Knowledge distillation (KD), mid–low-resolution images, remote sensing, ship detection, super-resolution (SR).

I. INTRODUCTION

MID–LOW-RESOLUTION (with a ground sample distance larger than 10 m/pixel) remote sensing ship detection is a valuable but challenging task since mid–low-resolution images cannot provide sufficient detail information as high-resolution (HR) images. To handle this challenge, recent methods [1]–[5] performed image super-resolution (SR) on mid–low-resolution images to recover their missing details and, thus, achieved improved detection performance. However, in these methods, ground-truth HR images are only used as labels to supervise the training of the SR module but not

leveraged in the detection stage, which limits the performance of mid–low-resolution ship detection.

Motivated by the recent advances of knowledge distillation (KD) [6]–[10], in this letter, we propose a feature distillation framework with a student network and a teacher network for mid–low-resolution ship detection. The student network is an SR-based detector containing an SR module (e.g., RDN [11]) and a detection module (e.g., *Faster-RCNN* [12]), while the teacher network only has a detection module, which is identical to the student one but takes the ground-truth HR images as inputs to generate HR feature representations. Then, the HR features produced by the teacher network are used as additional “ground-truth labels” to supervise the training of the student detection module through a feature distillation loss. In our framework, ground-truth HR images are not only used as labels to supervise the training of image SR but also used to provide informative knowledge to guide the training of ship detection, leading to significant performance improvement.

In summary, the contributions of our work are given as follows.

- 1) We propose a feature distillation framework for mid–low-resolution ship detection, where ground-truth HR images are used by both the SR module and the detection module of the student network.
- 2) We validate the effectiveness and generality of our framework by applying it to different detectors (e.g., *FCOS* [13], *Faster-RCNN* [12], *Mask-RCNN* [14], and *Cascade-RCNN* [15]). Experimental results demonstrate that our framework can introduce consistent performance improvements to these detectors.
- 3) We compare our method (with *Faster-RCNN* as the detection module) to several state-of-the-art detectors on three public datasets. Comparative results demonstrate the competitive performance of our method.

II. RELATED WORK

A. SR-Based Object Detection

Many works performed image SR to improve the detection performance on mid–low-resolution images. Dong *et al.* [1] proposed a second-order multiscale SR network and demonstrated the benefits of image SR to remote sensing object detection. Noh *et al.* [2] performed SR to enhance the features of a small region of interests (RoIs) and, thus, achieved improved detection performance on tiny objects. Courtrai *et al.* [3] proposed an object-focused SR approach and tailored it with a detection network to develop a generic detection framework. Wang *et al.* [4] proposed a feature-driven loss to reduce the task gap between SR and

Manuscript received June 29, 2021; revised August 16, 2021; accepted September 2, 2021. Date of publication September 15, 2021; date of current version December 20, 2021. This work was supported by the National Natural Science Foundation of China under Grant 62071474. (Corresponding author: Huanxin Zou.)

The authors are with the College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China (e-mail: heshitian19@nudt.edu.cn; hxzou2008@163.com; wangyingqian16@nudt.edu.cn; lirunlin14@nudt.edu.cn; chengfei08@nudt.edu.cn; cx2020@nudt.edu.cn; limeilin@nudt.edu.cn).

Digital Object Identifier 10.1109/LGRS.2021.3110404

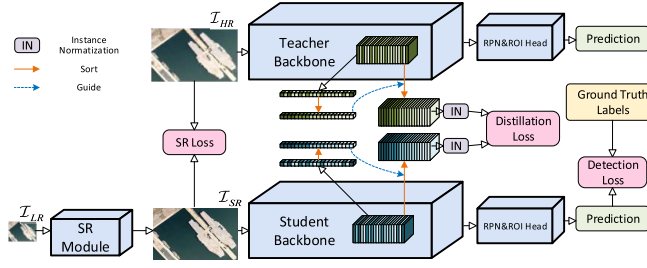


Fig. 1. Overview of our feature distillation framework.

detection. He *et al.* [5] fed both the super-resolved images and intermediate features to the detection network to achieve ship detection on mid–low-resolution images. However, in these methods, HR images were only used as ground-truth labels to supervise the training of image SR but not used by their detection modules, which hinders their detection performance.

B. KD

KD, which was originally proposed by Hinton *et al.* [6], has been widely used for model compression and acceleration. In the original settings of KD, a large teacher network and a lightweight student network take identical inputs, and the “knowledge” (e.g., response, feature, or relation) is “distilled” from the teacher to the student through various kinds of distillation losses [7]. Apart from this paradigm, the knowledge can be also obtained by feeding a more informative input to the teacher network. For example, Lee *et al.* [8] applied KD to image SR. In their method, HR images are fed to the teacher network to generate informative HR features to supervise the student network for image SR. Ge *et al.* [9] proposed a KD method to improve the performance of LR face recognition. In their method, the teacher network takes an HR face image as input to distill the most informative features to the student network (with an LR face image as input). In this letter, we follow the latter paradigm to use the teacher network to provide “ground-truth” HR features and design a feature distillation framework to enhance the detection performance of the student network.

III. METHODOLOGY

In this section, we first introduce the architecture of the student and the teacher networks and then describe the details of feature distillation.

A. Network Architecture

As shown in Fig. 1, our framework consists of a student network (with an SR module and a detection module) and a teacher network (with only a detection module).

In the student network, the SR module takes a mid–low-resolution image $I_{LR} \in \mathbb{R}^{H \times W \times 3}$ as its input to generate a super-resolved image $I_{SR} \in \mathbb{R}^{\alpha H \times \alpha W \times 3}$, where H and W represent the height and width of the input image, and α denotes the upsampling factor. As shown in Fig. 2, our SR module uses eight residual dense blocks (RDBs) [11] to extract residual feature F_{res} and integrates F_{res} with the initial feature F_0 through elementwise summation. The fused feature is then fed to a subpixel layer [16] (with $8 \times$ upsampling) to generate the residual prediction I_{res} , which is further added with the $8 \times$ bicubically upsampled input image to generate the final super-resolved image I_{SR} . Here, HR images are used as

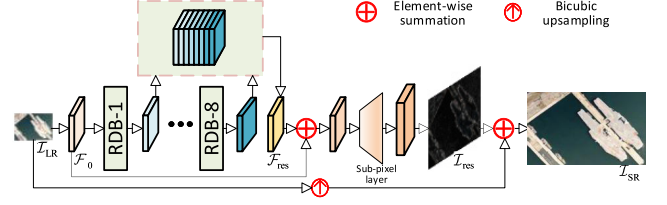


Fig. 2. Architecture of the SR module.

ground-truth labels to penalize incorrect super-resolved results and, thus, supervise the training of the SR module.

Given the super-resolved image generated from the SR module, a detection module is employed for ship detection. Since our framework is generic and can be applied to different detectors, we choose *Faster-RCNN* [12] as an example to introduce our framework. In the detection module, the super-resolved image is first fed to a backbone network (e.g., *ResNet101* [17]) for feature extraction. The extracted features are then fed to a region proposal network (RPN) to produce high-quality proposals. Both proposals and backbone features are fed to an ROI layer to generate final predictions.

In our framework, the teacher network only contains a detection module that is identical to the student one but directly takes the ground-truth HR image as its input. The HR features produced by the teacher network provide informative knowledge, which is then transferred to the student network through feature distillation. Details of the feature distillation will be introduced in Section III-B.

B. Feature Distillation

To transfer knowledge from the teacher to the student, we propose a feature distillation scheme to enforce the student to generate features as informative as the ground-truth one. To achieve this goal, a straightforward approach is to employ an L_1 or L_2 loss to minimize the distance between student feature and teacher feature. However, as illustrated in Figs. 3(a) and 4(a), feature maps of the teacher and the student differ significantly at the same channels. Consequently, directly performing pixelwise losses to their original features will make the student focus more on the discrepancy of channel distribution but ignore the real knowledge (discriminate spatial distribution) provided by the teacher, resulting in performance degradation (as demonstrated in Section IV-B).

Here, we propose a channel sorting approach to handle the channel discrepancy issue. Specifically, we first perform spatial global average pooling to calculate the channel intensity of the teacher and student features, respectively, and then sort these features along the channel dimension according to the calculated intensity values. It can be observed in Figs. 3(b) and 4(b) that the sorted features have similar channel distribution. Note that we further perform instance normalization [18] to further reduce the gap between the sorted teacher and student features. It is demonstrated in Section IV-B that our proposed channel sorting and normalization approach is effective and can introduce performance improvement to our framework.

C. Losses

The overall loss of our framework is defined as

$$L_{\text{overall}} = L_{\text{Det}} + \lambda_1 L_{\text{SR}} + \lambda_2 L_{\text{Distill}} \quad (1)$$

where L_{Det} , L_{SR} , and L_{Distill} represent detection loss, SR loss, and feature distillation loss, respectively. Here, L_{Det} is the

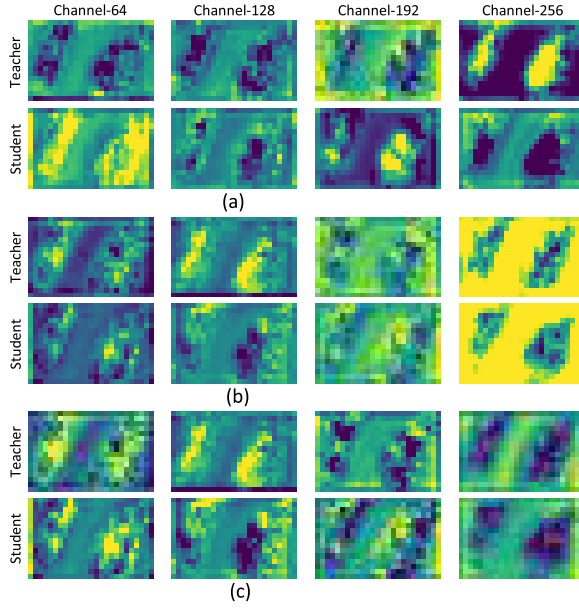


Fig. 3. Visualization of the distilled features before/after sorting and instance normalization. (a) Original features. (b) Sorted features. (c) Features after instance normalization.

detection loss, which is identical to the one used in *Faster-RCNN* [12]. L_{SR} denotes the SR loss and is the L_1 distance between the super-resolved image I_{SR} and the ground-truth image I_{HR} . That is

$$L_{SR} = \|I_{SR} - I_{HR}\|_1. \quad (2)$$

Moreover, $L_{distill}$ in (1) represents our proposed feature distillation loss. In our implementation, we chose the features in feature pyramid networks (FPNs) for distillation since FPN integrates multiscale features that contain beneficial information for detection [19]. Following [19], features of all the four levels from FPN are used for distillation, that is

$$L_{Distill} = \sum_{i=1}^4 \|\mathcal{H}(F_{teacher}^i) - \mathcal{H}(F_{student}^i)\|_2 \quad (3)$$

where $F_{teacher}^i$ and $F_{student}^i$ represent the features of the teacher and student networks at the i th level of their FPN, respectively. $\mathcal{H}$ denotes our proposed channel sorting and normalization operation.

IV. EXPERIMENT

In this section, we first introduce the datasets and implementation details and then conduct ablation studies to validate the effectiveness of our framework. Finally, we compare our method to several state-of-the-art methods on public datasets.

A. Datasets and Implementation Details

1) *Datasets*: We used the HRSC2016 [20], DOTA [21] and NWPU VHR-10 [22] datasets in our experiments.

1) HRSC2016 is an HR remote sensing dataset for ship detection, which contains 617 images for training and 438 images for validation. The image size in this dataset ranges from 300×300 pixels to 1500×900 pixels. We resized these images to 800×512 pixels as the ground-truth HR images.

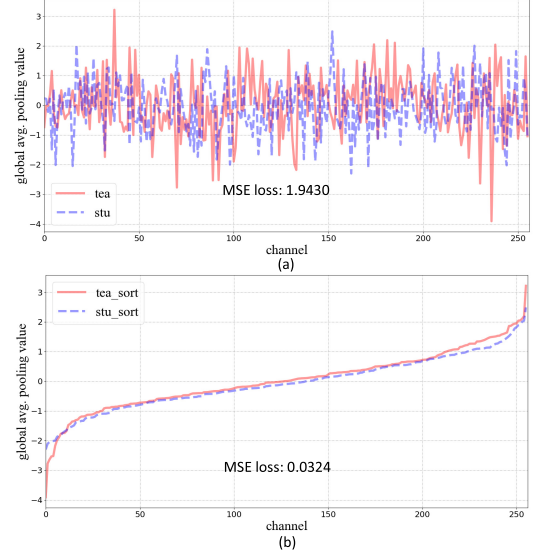


Fig. 4. Channel intensity distribution of the teacher and student features before and after the sorting operation. (a) Channel intensities of the original features. (b) Channel intensities of the sorted features.

- 2) DOTA is a large-scale dataset with 15 categories (including ships) for remote sensing object detection. We cropped the original images into patches of 512×512 pixels and chose the patches containing ship targets in our experiment. Finally, we generated 4163 images for training and 1411 images for validation.
- 3) NWPU VHR-10 is a multiclass object-detection dataset with ten categories. We cropped the original images into patches of 512×512 pixels and selected the ship patches for experiments. We totally obtained 249 images for training and 52 images for validation.

In our experiments, the above images (patches) were used as HR ground truths and bicubically downsampled by a factor of 8 to generate mid-low-resolution input images. We performed a series of data augmentation approaches, including random flipping, rotation, color transformation, brightness adjustment, and contrast transformation.

2) *Training Process*: We followed the training pipeline in [23] to train the student network using the mid-low-resolution images (generated by $8 \times$ bicubically downsampling on HR images) as inputs. Then, we trained the teacher network using the ground-truth HR images as inputs. Finally, we performed feature distillation using the pretrained student and teacher models. Note that, the teacher model was frozen during feature distillation, and only the parameters of the student network were updated. In each training stage, the learning rate was initially set to 5×10^{-3} and decreased by a factor of 0.1 for every ten epochs (totally 24 epochs).

3) *Evaluation Metrics*: We use the mean average precision (mAP) as the evaluation metric with the intersection over union (IoU) being set to 0.5 (i.e., mAP50) and 0.75 (i.e., mAP75).

B. Ablation Study

In this section, we conduct extensive ablation studies on the HRSC2016 dataset to validate the effectiveness of each component of our method.

TABLE I

DETECTION ACCURACY ACHIEVED BY DIFFERENT VARIANTS OF OUR PROPOSED METHOD. THE BEST RESULTS ARE IN BOLD FACES, AND THE SECOND BEST RESULTS ARE UNDERLINED

Model	FCOS		Faster-RCNN		Mask-RCNN		Cascade-RCNN	
	mAP50	mAP75	mAP50	mAP75	mAP50	mAP75	mAP50	mAP75
LR-Det	0.293	0.018	0.490	0.079	0.437	0.056	0.435	0.049
Bic-Det	0.653	0.259	0.820	0.626	0.835	0.654	0.831	0.682
SR-Det	0.713	0.328	0.863	0.712	0.850	0.713	0.862	0.750
Student	0.800	0.511	0.875	0.762	0.864	0.740	0.886	0.784
Teacher	0.818	0.563	0.910	0.769	0.911	0.783	0.907	0.804

TABLE II

COMPARATIVE RESULTS ACHIEVED BY OUR METHOD WITH/WITHOUT SORTING AND NORMALIZATION

$L_{Distill}$	FCOS		Faster-RCNN		Mask-RCNN		Cascade-RCNN	
	mAP50	mAP75	mAP50	mAP75	mAP50	mAP75	mAP50	mAP75
Vanilla L2	0.790	0.483	0.877	0.749	0.864	0.716	0.875	0.770
Proposed	0.800	0.511	0.875	0.762	0.864	0.740	0.886	0.784

1) *SR for Ship Detection*: To demonstrate the benefits introduced by image SR to ship detection, we first removed the SR module in the student network and fed the mid-low-resolution images and their bicubically upsampled versions to the detection module, respectively. Then, we used our SR module for upsampling and fed the super-resolved images to the detection module. As shown in Table I, directly using the mid-low-resolution images (i.e., LR-Det) achieves very low detection accuracy. When the bicubic upsampling is introduced (i.e., Bic-Det), the detection performance is significantly improved. Moreover, using our SR module (i.e., SR-Det) can further enhance the detection performance, which demonstrates the benefits of SR to ship detection.

2) *Feature Distillation*: To validate the effectiveness of feature distillation, we compare the detection performance of the student network with/without performing feature distillation. As shown in Table I, by performing feature distillation, the student network can further achieve a 0.03–0.08 improvement over SR-Det in terms of mAP75. That is because the teacher network takes HR images as its inputs to provide ground-truth features. With the additional features for supervision, the student can better learn to extract informative features from the super-resolved images for ship detection.

3) *Channel Sorting and Normalization*: We investigate the effectiveness of the channel sorting and feature normalization approach by directly applying an L_2 loss to the teacher and student features. It can be observed in Table II that the detection accuracy suffers a minor decrease if the channel sorting and normalization are canceled. As demonstrated in Section III-B, the proposed channel sorting and normalization can reduce the gap between teacher and student features, which can make the student network focus more on the spatial discrepancy and facilitate the knowledge transfer.

4) *Positions for Distillation*: Since the FPN module has four levels of features with different resolutions, we compare the performance of our framework with different features being used for distillation. As shown in Table III, these variants achieve comparable detection performance, which demonstrates that our framework is relatively insensitive to the choice of features for distillation. Based on the results, we chose all the four-level features for distillation, as formulated in (3).

5) *Weights of Losses*: We compare the performance of our method with different settings of λ_1 and λ_2 in (1). As shown in Table IV, a proper value of the weights of losses can

TABLE III

COMPARATIVE RESULTS ACHIEVED BY OUR METHOD USING DIFFERENT FEATURES FROM FPN FOR DISTILLATION

Layers				Detection Performance	
04	03	02	01	mAP50	mAP75
✓				0.878	0.759
	✓			0.878	0.758
		✓		0.879	0.756
			✓	0.878	0.752
✓	✓	✓	✓	0.875	0.762

TABLE IV

RESULTS ACHIEVED BY OUR METHOD WITH DIFFERENT SETTINGS OF λ_1 AND λ_2 . THE BEST RESULTS ARE IN BOLD FACES, AND THE SECOND BEST RESULTS ARE UNDERLINED

λ_2 ($\lambda_1=1$)	0	0.0001	0.001	0.01	0.1	1	10
mAP50	0.863	0.879	<u>0.875</u>	0.868	<u>0.875</u>	0.859	NAN
mAP75	0.712	0.752	0.762	0.758	0.740	0.708	NAN
λ_1 ($\lambda_2=0.001$)	0	0.001	0.01	0.1	1	10	100
mAP50	0.866	0.869	0.873	0.877	0.875	0.875	0.871
mAP75	0.725	0.731	0.732	0.739	0.762	<u>0.742</u>	0.723

TABLE V

QUANTITATIVE RESULTS ACHIEVED BY DIFFERENT METHODS (BASED ON RESNET101) ON THE HRSC2016, DOTA, AND NWPU VHR-10 DATASETS. HERE, WE ADOPT THE *Faster-RCNN* DETECTOR AS OUR DETECTION MODULE. THE BEST RESULTS ARE IN BOLD FACES

Method	HRSC2016		DOTA		NWPU VHR-10		#Params. (MB)	Time (ms)
	mAP50	mAP75	mAP50	mAP75	mAP50	mAP75		
GFL [27]	0.826	0.632	0.575	0.181	0.856	0.488	411.2	46.95
Reppoints [26]	0.792	0.453	0.592	0.161	0.875	0.563	331.8	51.28
HTC [24]	0.634	0.579	0.668	0.296	0.863	0.568	791.3	104.2
DetectoRS [25]	0.860	0.735	0.677	0.311	0.808	0.604	1337	158.7
Faster-RCNN [12]	0.820	0.626	0.651	0.233	0.853	0.549	485.6	48.31
Ours	0.875	0.762	0.709	0.340	0.875	0.604	503.3	56.50

*Inference time is averaged on the HRSC2016 dataset with an input mid-low resolution image of size 100×64 pixels.

make our method achieve better performance. Based on the comparative results, we empirically set $\lambda_1 = 1$, $\lambda_2 = 0.001$ in our experiments.

C. Comparison to State-of-the-Art Methods

We apply our framework to *Faster-RCNN* and compare our method with four state-of-the-art detection methods, including *HTC* [24], *DetectoRS* [25], *Reppoints* [26], and *GFL* [27]. We use the bicubically upsampled images as the inputs of the compared methods to ensure that the input size of different detectors is identical.

1) *Quantitative Results*: Quantitative results are shown in Table V. It can be observed that our method outperforms *Reppoints*, *GFL*, *HTC*, and *DetectoRS* on all the three datasets and achieves significant improvements over the original *Faster-RCNN*. Compared with *DetectoRS*, our method achieves a better performance with much fewer parameters and less inference time. Note that, although *HTC* and *DetectoRS* are also developed on *Faster-RCNN*, these two methods are less competitive due to the missing details in the input bicubically upsampled images. In contrast, by performing image SR and feature distillation, our method can well handle this problem and achieves state-of-the-art detection accuracy.

2) *Qualitative Results*: Fig. 5 shows the qualitative results achieved by different methods on three example scenes. It can be observed that our method generates more accurate bounding boxes with fewer misdetection and false alarms. For example, the two ships in Scene A are similar in appearance and park closely. *GFL*, *Reppoints*, and *HTC* cannot separate these two

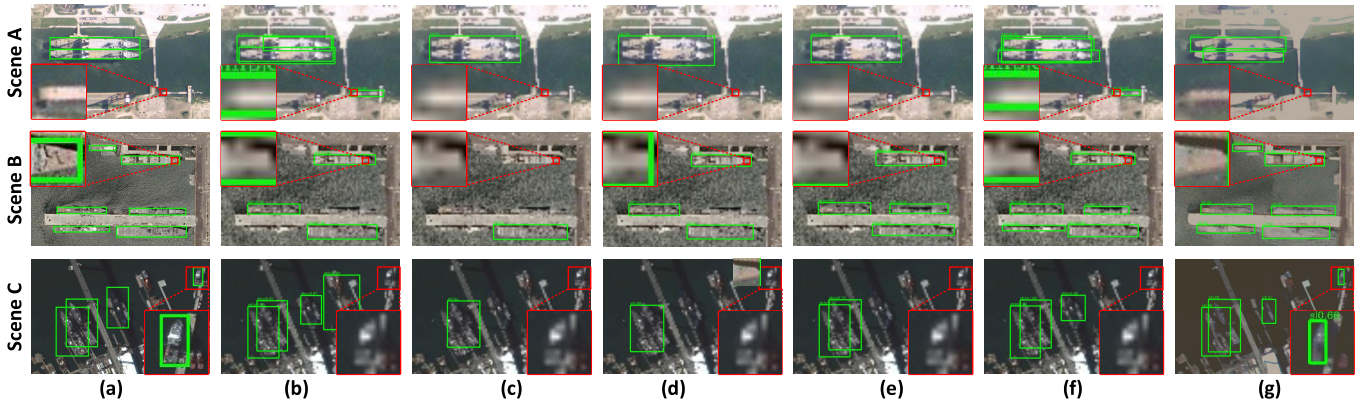


Fig. 5. Qualitative results achieved by different methods on three example scenes. “HR” denotes HR images, and “GT” represents ground-truth labels. We use green bounding boxes to mark the detection results and red boxes to highlight zoomed-in regions. (a) HR + GT. (b) Faster-RCNN. (c) GFL. (d) Reppoints. (e) HTC. (f) DetectoRS. (g) Ours.

ships, and *Faster-RCNN* and *DetectoRS* generate false alarms in the right-bottom regions. In contrast, our method does not suffer from these problems and achieves accurate ship detection. The results in Scenes B and C also demonstrate the superior performance of our method in handling mid-low-resolution ship detection tasks.

V. CONCLUSION

In this letter, we propose a feature distillation framework to enhance the mid-low-resolution ship detection. In our framework, HR images are not only used as ground-truth labels to train the SR module but also provide “ground-truth” features to train the detection module, which significantly enhances the detection performance of the student network. Extensive ablation studies have demonstrated the effectiveness of our framework with different detection modules. Comparative results on three public datasets have demonstrated that our framework can further improve the detection performance of SR-based detectors for mid-low-resolution ship detection and can achieve higher detection accuracy compared to state-of-the-art detection methods.

REFERENCES

- [1] X. Dong, L. Wang, X. Sun, X. Jia, L. Gao, and B. Zhang, “Remote sensing image super-resolution using second-order multi-scale networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3473–3485, Apr. 2020.
- [2] J. Noh, W. Bae, W. Lee, J. Seo, and G. Kim, “Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 9725–9734.
- [3] L. Courtrai, M.-T. Pham, and S. Lefèvre, “Small object detection in remote sensing images based on super-resolution with auxiliary generative adversarial networks,” *Remote Sens.*, vol. 12, no. 19, p. 3152, Sep. 2020.
- [4] B. Wang, T. Lu, and Y. Zhang, “Feature-driven super-resolution for object detection,” in *Proc. CRC*, 2020, pp. 211–215.
- [5] S. He, H. Zou, Y. Wang, R. Li, and F. Cheng, “ShipSRDet: An end-to-end remote sensing ship detector using super-resolved feature representation,” in *Proc. IGARSS*, 2021, pp. 1–4.
- [6] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in *Proc. NIPS*, Mar. 2015, pp. 1–9.
- [7] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge distillation: A survey,” *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.
- [8] W. Lee, J. Lee, D. Kim, and B. Ham, “Learning with privileged information for efficient image super-resolution,” in *Proc. ECCV*, Aug. 2020, pp. 465–482.
- [9] S. Ge, S. Zhao, C. Li, and J. Li, “Low-resolution face recognition in the wild via selective knowledge distillation,” *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2051–2062, Nov. 2018.
- [10] Z. Xiao, X. Fu, J. Huang, Z. Cheng, and Z. Xiong, “Space-time distillation for video super-resolution,” in *Proc. CVPR*, 2021, pp. 2113–2122.
- [11] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in *Proc. CVPR*, 2018, pp. 2472–2481.
- [12] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
- [13] Z. Tian, C. Shen, H. Chen, and T. He, “FCOS: Fully convolutional one-stage object detection,” in *Proc. ICCV*, 2019, pp. 9627–9636.
- [14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in *Proc. ICCV*, 2017, pp. 2961–2969.
- [15] Z. Cai and N. Vasconcelos, “Cascade R-CNN: High quality object detection and instance segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1483–1498, May 2021.
- [16] W. Shi *et al.*, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *Proc. CVPR*, 2016, pp. 1874–1883.
- [17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. CVPR*, 2016, pp. 770–778.
- [18] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Instance normalization: The missing ingredient for fast stylization,” 2016, *arXiv:1607.08022*. [Online]. Available: <http://arxiv.org/abs/1607.08022>
- [19] X. Dai *et al.*, “General instance distillation for object detection,” in *Proc. CVPR*, 2021, pp. 7842–7851.
- [20] Z. Liu, L. Yuan, L. Weng, and Y. Yang, “A high resolution optical satellite image dataset for ship recognition and some new baselines,” in *Proc. ICPR*, vol. 2, 2017, pp. 324–331.
- [21] G.-S. Xia *et al.*, “DOTA: A large-scale dataset for object detection in aerial images,” in *Proc. CVPR*, 2018, pp. 3974–3983.
- [22] G. Cheng, J. Han, P. Zhou, and L. Guo, “Multi-class geospatial object detection and geographic image classification based on collection of part detectors,” *ISPRS J. Photogramm. Remote Sens.*, vol. 98, pp. 119–132, Dec. 2014.
- [23] S. He, H. Zou, Y. Wang, R. Li, F. Cheng, and X. Cao, “Mid-low resolution remote sensing ship detection using super-resolved feature representation,” Tech. Rep. 202108.0337, 2021.
- [24] K. Chen *et al.*, “Hybrid task cascade for instance segmentation,” in *Proc. CVPR*, 2019, pp. 4974–4983.
- [25] S. Qiao, L.-C. Chen, and A. Yuille, “Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution,” in *Proc. CVPR*, 2021, pp. 10213–10224.
- [26] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, “RepPoints: Point set representation for object detection,” in *Proc. ICCV*, 2019, pp. 9657–9666.
- [27] X. Li *et al.*, “Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection,” in *Proc. Comput. Vis. Pattern Recognit.*, 2020, pp. 1–11.