EasyChair Preprint

№ 2661

Semi-supervised Synthetic-to-Real Domain Adaptation for Fine-grained Naval Ship Image Classification

Yoonhyung Kim, Hyeonjin Jang, Sangtae Park, Jiwon Lee and Changick Kim

February 14, 2020

Semi-supervised Synthetic-to-Real Domain Adaptation for Fine-grained Naval Ship Image Classification*

Yoonhyung Kim[1][0000−0002−5608−8473], Hyeonjin Jang[2], Sangtae Park[2], Jiwon Lee[3], and Changick Kim[1][0000−0001−9323−8488]

[1] Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea
{yhkim1127,changick}@kaist.ac.kr
[2] SafeTechResearch Inc., Daejeon, Republic of Korea
{namuland, stpark}@strkorea.co.kr
[3] Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea
ez1005@etri.re.kr

Abstract. In this paper, we propose a deep learning-based approach for fine-grained naval ship image classification. To this end, we tackle following two major challenges. First, to overcome the lack of the amount of training images in the target (i.e., real) domain, we generate a large number of synthetic naval ship images by using a simulation program which is specifically designed for our task. Second, to relieve performance degradation caused by the disparity between the synthetic and the real domains, we propose a novel regularization loss, named cross-domain triplet loss. Experimental results show that both the synthetic images and the proposed cross-domain triplet loss are essential to achieve the state-of-the-art performance for fine-grained naval ship image classification.

Keywords: Domain adaptation · Fine-grained image classification · Deep learning.

1 Introduction

Along with the rapid development of deep learning-based computer vision technologies, it has become a worldwide trend to develop intelligent systems that can conduct various tasks which require recognition ability. One of the primary roles of those systems is to support human operators by substituting repetitive and tedious tasks. As a representative example, camera-based automatic surveillance

* This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2019-0-00524, Development of precise content identification technology based on relationship analysis for maritime vessel/structure).

systems can be used to assist human observers by detecting and identifying objects in surveilling regions. Based on this perspective, we aim to develop such an automatic system that surveils maritime areas, and the topic of this paper is focused on identifying objects (i.e., naval ship classification) on the assumption that the object locations are given.

There are two major challenges to develop a deep learning-based naval ship classifier. First, a large-scale dataset is required to train deep neural networks, but mining a large number of real naval ship images is realistically infeasible due to the security issue. To overcome this limitation, we propose to generate a large number of synthetic naval ship images and adopt them for training. The second challenge is to maximize the classification performance by relieving the domain disparity between source (synthetic) and target (real) domains. To this end, we apply the minimax entropy method [13], which is one of the state-of-the-art semi-supervised domain adaptation approaches. In addition, to further enhance the classification performance, we propose a novel regularization loss, named cross-domain triplet loss and embed the loss for the domain adaptive learning process. Experimental results demonstrate that both the synthetic images and the proposed triplet loss serve as key components to achieve the state-of-the-art performance for fine-grained naval ship classification.

The rest of this paper is organized as follows. In Section 2, we introduce our naval ship datasets. In Section 3, we explain the details of domain adaptive learning process along with the proposed cross-domain triplet loss. Experimental results and analysis are presented in Section 4 and concluding remarks are given in Section 5.

2 Synthetic & Real Naval Ship Image Dataset

In this section, we introduce our naval ship image dataset. As mentioned earlier, the naval ship dataset is composed of two domains, i.e., synthetic and real domains. Both the synthetic and the real domain images share the same set of classes. In Table 1, the details of the naval ship classes are given. For the eight kinds of naval ships, we train and validate fine-grained classifiers. Note that, in our paper, we call this task as "fine-grained classification" because all the images belong to the same class "naval ship" and our goal is to distinguish the slight differences among the eight sub-classes. A set of synthetic and real image samples for the eight kinds of naval ships is illustrated in Fig. 1.

2.1 Synthetic Naval Ship Images

The process of generating synthetic naval ship images consists of following four stages. First, various specifications of each naval ship are collected. Note that, in this stage, only publicly available data (e.g., elementary specifications and rough shapes of equipment for each naval ship) are collected by means of ordinary web searching. Second, based on the specifications, 3D models of each naval ship are rendered by using a 3D simulation program. As the third stage, texture

Table 1. Specifications of the Korean naval ships that are used for our experiments.

No.	ID	Full name	Naval ship class	Displacement (tons)
1	FF	Frigate	Frigate	1500
2	FFG	Frigate Guided-missile	Frigate	2800˜3300
3	KDX-1	Korea Destroyer eXperimental-1	Destroyer	3200
4	KDX-2	Korea Destroyer eXperimental-2	Destroyer	4500
5	PCC	Patrol Combat Corvette	Patrol Corvette	950
6	PKG	Patrol Killer Guided-missile	Patrol Vessel	440
7	PKM	Patrol Killer Medium	Patrol Killer	150
8	YUB	Yard Utility Boat	Patrol Boat	55

mapping and refinements are conducted to make the 3D models more realistic and aesthetic. The final stage is to generate a large number of naval ship images for various viewpoints by capturing the screen of the 3D simulation program. Through this process, we obtained 17,811 synthetic naval ship images. The class-wise quantities of synthetic images are 1,617, 1,840, 2,178, 2,397, 1,543, 1,744, 3,546, 2,946, respectively (in the order of the numbers in Table 1).

2.2 Real Naval Ship Images

By means of web image searching, we collected 124 naval ship images. The real images are manually captured so that each naval ship's identity can be clearly represented. The number of samples for each class ranges from 10 to 20. By comparing the naval ship images in Fig. 1, we can intuitively recognize the domain disparity between the synthetic and the real images. The challenge here is to make deep neural networks be robust to domain shifts and fully exploit the whole training data for classification. This can be done by applying a domain adaptive training scheme which is explained in the next section.

3 Semi-supervised Domain Adaptation for Fine-grained Naval Ship Classification

3.1 Related Work

Domain adaptation is one of the transfer learning schemes for deep neural networks. Specifically, the aim of domain adaptation techniques is to transfer the knowledge acquired by labeled data in a source domain to a target domain under the condition that the amount of labeled data in the target domain is very scarce. Since mining a large number of labeled data to train deep neural networks is often very expensive and time-consuming, domain adaptive learning techniques can be utilized to resolve this practical issue. Another setting for domain adaptation in our consideration is that source and target domains share the same set of classes.

Fig. 1. Visual comparison of the synthetic (left column) and the real (right column) naval ship images. The class names are overlaid on the upper left side of the images.

Generally, existing domain adaptation methods for image classification are categorized into two major approaches, i.e., unsupervised and semi-supervised approaches. The first one configures the case that only unlabeled images are given for target domain while labeled images in source domain are accessible. As an early work for unsupervised domain adaptation, Ganin and Lempitsky [3] propose an adversarial learning approach by establishing a minimax game between a feature generator and a domain classifier. By assigning a domain label for each domain (e.g., 0 and 1 for source and target domains, respectively), the feature extractor is trained to deceive the domain classifier by generating domain-invariant features. This process is implemented via the gradient reversal layer (GRL). To further enhance the discriminative ability of classifiers, Long et al. [10] propose an conditional adversarial learning strategy by using classifier predictions as auxiliary clues. There are several methods that utilize generative adversarial networks (GANs) for domain adaptation. In [17], Volpi et al. propose a data augmentation approach by using a GAN-based approach. Hu et al. [6] propose to use GANs for feature alignment, and they devise duplex discriminators to further enhance discriminative powers of features. The limitation of the GAN-based approaches is that those methods require a large number of images to train GANs, which is infeasible in our case. Motivated by the observation that feature samples which are located near the decision boundary of a classifier often result in misclassification, Saito et al. [14] propose to alternately maximize and minimize the consensus of two classifiers by updating the feature generator and the classifiers by turns. There are many other approaches for unsupervised domain adaptation such as an attention-guided method [9], using a pseudo labeling strategy [1], self ensembling-based method [2], applying style transfer methods for pixel-level domain transformation [4], adopting graph neural networks [11].

The second approach is semi-supervised domain adaptation that a few amount of labeled data in a target domain are given. Although the above-mentioned unsupervised methods can be adopted for semi-supervised domain adaptation, several methods specialized for semi-supervised settings are also proposed recently. Saito et al. [13] indicate that several unsupervised methods occasionally perform even worse under semi-supervised settings, and they propose the minimax entropy-based method for semi-supervised domain adaptation. The key idea is to minimize the distance between the class prototypes and neighboring unlabeled target samples to extract discriminative features. To this end, they alternately maximize and minimize the conditional entropy of target data with respect to the classifier and the feature generator, respectively. This minimax entropy-based method achieves the state-of-the-art performance for semi-supervised domain adaptation and we adopt this approach as the baseline for fine-grained naval ship classification.

3.2 Semi-supervised Domain Adaptation via Minimax Entropy and Cross-domain Triplet Loss

Our goal is to train a naval ship classification model by using image data in the source (synthetic) and the target (real) domains. In the source domain, we

are given source images and the corresponding labels $\mathcal{D}_s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$. In the target domain, unlabeled images $\mathcal{D}_u = \{(\mathbf{x}_i^u)\}_{i=1}^{n_u}$ and a small fraction of labeled images $\mathcal{D}_t = \{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{n_t}$ are given. Using the three image sets, we train a classification model on $\mathcal{D}_s$, $\mathcal{D}_u$, $\mathcal{D}_t$ and we test on $\mathcal{D}_u$. The classification model is composed of a feature extractor F and a classifier C. For an input image $\mathbf{x}$, the output prediction of the model is denoted as $\mathbf{p}(\mathbf{x})$.

As indicated earlier, we adopt the minimax entropy-based training method [13] as the baseline. Two objective functions are used for training and the first one is the standard cross-entropy loss to train F and C:

$$\mathcal{L}_{sup} = \mathbb{E}_{(\mathbf{x},y)\in\mathcal{D}_s,\mathcal{D}_t}\mathcal{L}_{ce}(\mathbf{p}(\mathbf{x}), y). \tag{1}$$

With the guidance of the above loss function, the model is trained to discriminate naval ship images with respect to the synthetic source images and the real target images. However, since the number of labeled images in the target domain is much smaller than that of the source domain, the model is inclined to be biased to the source domain. To overcome this limitation, the entropy is introduced as the second objective function which is defined as follows [13]:

$$H = -\mathbb{E}_{\mathbf{x}\in\mathcal{D}_u}\sum_{i=1}^{K} p(y=i|\mathbf{x})\ \log\ p(y=i|\mathbf{x}), \tag{2}$$

where K indicates the number of naval ship classes, which is set to eight in our experiments. In (2), $p(y=i|\mathbf{x})$ is the probability of an output prediction for the ith class. The entropy function is maximized for the uniform output distribution and is minimized for the one-hot output distribution. The key motivation of the minimax entropy method [13] consists of the following two components. First, by increasing the entropy with respect to the feature generator, the feature distributions of two different domains get closer resolving the domain disparity. On the other side, by decreasing the entropy with respect to the classifier, the classifier is induced to produce more discriminative outputs. By incorporating these motivations, the adversarial objective function is defined as follows[13]:

$$\hat{W_F} = \underset{W_F}{\operatorname{argmin}}\ \mathcal{L}_{sup} + \lambda_{ent}H, \tag{3}$$

$$\hat{W_C} = \underset{W_C}{\operatorname{argmin}}\ \mathcal{L}_{sup} - \lambda_{ent}H, \tag{4}$$

where $\hat{W_F}$ and $\hat{W_C}$ indicate the weight parameters of the feature extractor and the classifier, respectively. The constant λ_{ent} in (4) is the balancing factor.

In our experimental setting, not only the amount of labeled target images but also that of unlabeled target images are very scarce. As a result, the effectiveness of the minimax entropy-based method is restricted in our case. Based on this observation, we propose to apply a novel loss function named cross-domain triplet loss to further enhance the domain alignment of the feature distribution. Our

proposed cross-domain triplet loss is defined as follows:

$$\mathcal{L}_{tri} = \mathbb{E}_{\mathbf{x}_s \in \mathcal{D}_s, \mathbf{x}_t \in \mathcal{D}_t} \max\left(\left\| f(\mathbf{x}_{t_a}) - f(\mathbf{x}_{s_p}) \right\|_2^2 - \| f(\mathbf{x}_{t_a}) - f(\mathbf{x}_{s_n}) \|_2^2 + \alpha_{tri}, 0 \right), \tag{5}$$

where $f(\cdot)$ indicates the function of feature extraction by F. In (5), $\mathbf{x}_{t_a}$ denotes an anchor image in the target domain, $\mathbf{x}_{s_p}$ is a source image whose class is identical to that of the anchor image, and $\mathbf{x}_{s_n}$ is a source image whose class is different from that of the anchor image. For each anchor image, two feature distances are measured via the l_2 norm. Based on the feature distances, the triplet loss encourages the distance between a feature pair with the same class be smaller, and vice versa. In this way, the feature extractor F is encouraged to produce more discriminative features across the two domains. The constant α_{tri} in (5) is a marginal threshold for the triple loss.

By incorporating the proposed triple loss, the final adversarial objective function is established as follows:

$$\hat{W_F} = \underset{W_F}{\operatorname{argmin}} \ \mathcal{L}_{sup} + \lambda_{tri}\mathcal{L}_{tri} + \lambda_{ent}H, \tag{6}$$

$$\hat{W_C} = \underset{W_C}{\operatorname{argmin}} \ \mathcal{L}_{sup} + \lambda_{tri}\mathcal{L}_{tri} - \lambda_{ent}H. \tag{7}$$

To conduct the adversarial learning with the above objective, a gradient reversal layer (GRL) [3] is inserted between the feature extractor and the classifier. The details of experimental settings and results are given in the next section.

4 Experiments

4.1 Experimental Setups

To train the networks, we used 17,811 synthetic naval ship images as the source image set $\mathcal{D}_s$. For the target domain, we randomly selected 5 images for each class (thus, 40 images in total) for the labeled image set $\mathcal{D}_t$ and used the other 84 images for the unlabeled image set $\mathcal{D}_u$. All experiments in this paper are implemented in PyTorch [12]. To demonstrate the consistency of performance improvements, we employ AlexNet [8], VGGNet [16], ResNet [5], and DenseNet [7] for the experiments. The batch sizes for $\mathcal{D}_s, \mathcal{D}_t, \mathcal{D}_u$ are set to 16, 4, 16, respectively. The balancing factors λ_{tri} and λ_{ent} are set to 0.1. The marginal threshold α_{tri} is set to 2.0 for AlexNet, 1.0 for VGGNet, and 3.0 for ResNet and DenseNet. We followed [13] for all the other training setups such as feature normalizations, learning rates, and data augmentations (horizontal flipping and random cropping).

4.2 Experimental Results and Analysis

The comparative evaluation results are given in Table 2. To validate the effectiveness of our proposed method, we compare with four other training approaches.

Table 2. Performance comparison in terms of classification accuracy (%) for fine-grained naval ship image classification.

Method	Baseline architecture					
	AlexNet	VGGNet	ResNet-18	ResNet-50	DenseNet-121	DenseNet-161
S only	20.64	22.62	23.81	22.62	20.64	24.60
T only	53.17	58.12	44.41	48.41	41.67	53.17
S+T	51.98	59.52	58.30	62.32	73.91	73.43
MME	55.16	66.00	61.11	61.51	76.59	74.21
Proposed	**63.49**	**77.78**	**65.88**	**63.49**	**78.97**	**76.59**

In Table 2, 'S only' and 'T only' indicate the training strategies that exploit labeled source images and labeled target images, respectively. 'S+T' stands for training by using both labeled source and target images. 'MME' means minimax entropy-based training method [13]. All numerical results in Table 2 are the average score obtained by three times of training.

Various analysis can be found by the results in Table 2. First, training by using the images from both domains (S+T) generally leads to the better performance than using either the synthetic (S only) or the real (T only) images only. This indicates that the synthetic images obviously contribute to achieve the better performance on the real domain. Second, the domain adaptive learning method with minimiax entropy leads to the better performance than the non-adaptive training strategy. By using our proposed method, the accuracies are further increased achieving the best performances for all cases. In particular, applying the triplet loss for our task leads to a large increasement of performance for simpler networks such as AlexNet and VGGNet. Therefore, our proposed triplet loss can be used to obtain the optimal performance upon the situation that computational resources are restricted. Finally, the accuracies are not always proportional to the depth of the networks. Since we used very small number of target images, the networks are prone to be overfitted. We anticipate that the performance would be increased for deeper networks if the amount of target images becomes larger.

To further analyze the results, we generated visualization images by using the gradient-weighted class activation mapping (Grad-CAM) algorithm [15]. A Grad-CAM result is a heatmap which highlights image regions that are paid attentions by a classification network. Thus, we can deduce the decision process of the network by observing Grad-CAM results. In Fig. 2, we can see that the classification network generally concentrates on the most discriminative parts of naval ships such as radar antennas (FF, PCC), canons (FFG), and other apparatuses. On the other hand, as illustrated in Fig. 3, the classification network fails to correctly recognize the naval ships when an image is blurred by a smoking effect (FFG) or when apparatuses or decoration patterns are very similar to those of other classes (KDX-1, PKM).

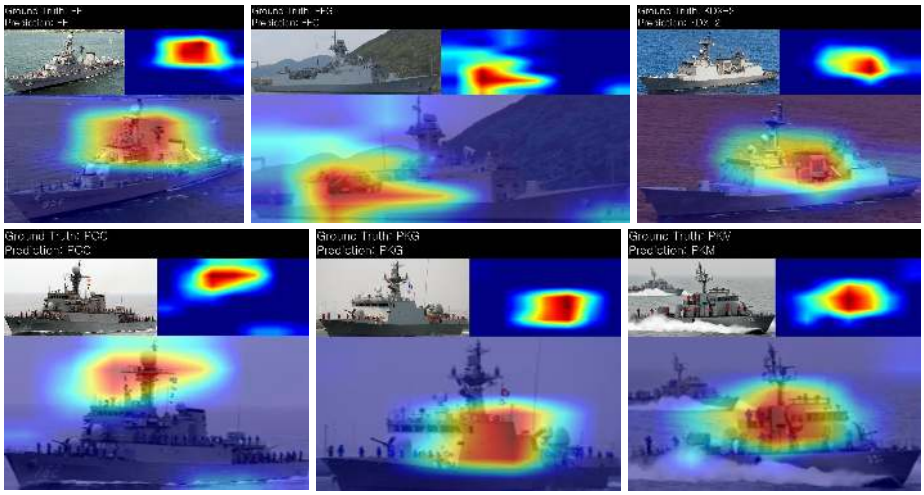

Fig. 2. Grad-CAM[15] visualization of correctly classified samples. The classification network is DenseNet-161 [7].

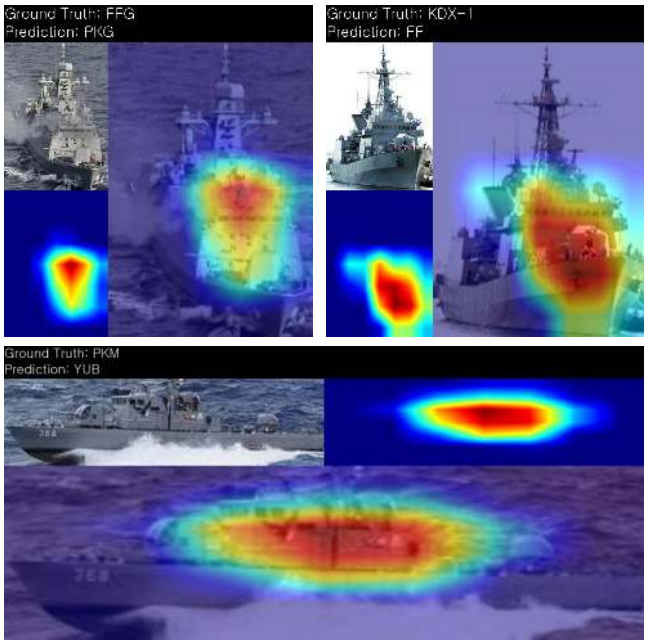

Fig. 3. Grad-CAM[15] visualization of failure cases. The classification network is DenseNet-161 [7].

5 Conclusion

In this paper, we have proposed a deep learning-based approach for fine-grained naval ship image classification. The major contribution of our work consists of the following two components. First, we produced a large number of synthetic naval ship images and utilized them for training deep neural networks. Second, we proposed a novel cross-domain triplet loss to align features of two distinct domains. By means of extensive comparative evaluations, the effectiveness of using synthetic images and the triplet loss are demonstrated. We expect that our work would be a useful and practical benchmark for researchers in computer vision fields.

References

1. Choi, J., Jeong, M., Kim, T., Kim, C.: Pseudo-labeling curriculum for unsupervised domain adaptation. British Machine Vision Conference (BMVC) (2019)
2. French, G., Mackiewicz, M., Fisher, M.: Self ensembling for visual domain adaptation. arXiv preprint arXiv:1706.05208 (2017)
3. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. International Conference on Machine Learning (ICML) (2015)
4. Gong, R., Li, W., Chen, Y., Gool, L.: Dlow: Domain flow for adaptation and generalization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 770–778 (2016)
6. Hu, L., Kan, M., Shan, S., Chen, X.: Duplex generative adversarial network for unsupervised domain adaptation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
7. Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., Keutzer, K.: Densenet: Implementing efficient convnet descriptor pyramids. arXiv preprint arXiv:1404.1869 (2014)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Neural Information Processing Systems (NeurIPS) (2012)
9. Kurmi, V., Kumar, S., Namboodiri, V.: Attending to discriminative certainty for domain adaptation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
10. Long, M., Cao, Z., Wang, J., Jordan, M.: Conditional adversarial domain adaptation. Neural Information Processing Systems (NeurIPS) (2018)
11. Ma, X., Zhang, T., Xu, C.: Gcan: Graph convolutional adversarial network for unsupervised domain adaptation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
12. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
13. Saito, K., Kim, D., Sclaroff, S., Darrell, T., Saenko, K.: Semi-supervised domain adaptation via minimax entropy. Proceedings of the IEEE Conference on International Conference on Computer Vision (ICCV) (2019)

14. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
15. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV) pp. 618–626 (2017)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
17. Volpi, R., Morerio, P., Savarese, S., Murino, V.: Adversarial feature augmentation for unsupervised domain adaptation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)