DYFUSION: DYNAMIC IR/RGB FUSION FOR MARITIME VESSEL RECOGNITION

Cassio E. Santos Jr. and Bir Bhanu

Center for Research in Intelligent Systems
University of California, Riverside, CA 92521, USA
celias@engr.ucr.edu, bhanu@cris.ucr.edu

ABSTRACT

We propose a novel multi-sensor data fusion approach called DyFusion for maritime vessel recognition using long-wave infrared and visible images. DyFusion consists of a decision-level fusion of convolutional networks using a probabilistic model that can adapt to changes in the scene. The probabilistic model avails of contextual clues from each sensor decision pipeline to maximize accuracy and to update probabilities given to each sensor pipeline. Additional sensors are simulated by applying simple transformations on visible images. Evaluation is presented on the VAIS dataset, demonstrating the effectiveness and robustness of DyFusion with a reliable accuracy of up to 88% in hard scenarios.

Index Terms— maritime vessel recognition, sensor fusion, convolutional neural networks, probabilistic models

1. INTRODUCTION

The use of multiple sensors provides better results in many tasks such as face [1], activity [2] and maritime vessel recognition [3]. If a set of homogeneous sensors are used, the uncertainty about the measurements is reduced due to the law of large numbers. If a set of heterogeneous sensors are used, complementary data from each sensor improves the prediction of hidden variables by conditioning the prediction to the independent observations from each sensor. Unfortunately, the best way to combine multi-sensor data for a task is often unintuitive and there is no ideal approach for each circumstance [4]. Hence, many multi-sensor data fusion[1] approaches have been published over the years.

The Joint Directors of Laboratories (JDL) Data Fusion Working Group defines a terminology for six subprocesses used in sensor fusion. We focus on L1 and L4: recognizing and adapting the fusion model to changes in the scene. A technical description of JDL subprocesses can be found in [4]. We also focus on the long-wave infrared (LWIR[2]) and visible images for maritime vessel recognition when evaluating DyFusion. Vessel recognition is an important task for safety and regulation enforcement, specially because many products are transported via ocean [3]. IR images provide invariance to changes in lighting resulted from atmospheric conditions and from images acquired in various times during the day. For instance heat emitted by the boat internal combustion motor improves its visibility in the IR spectrum during night.

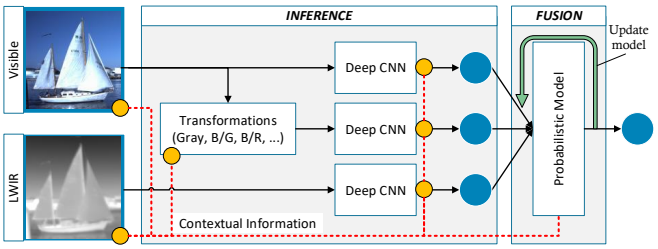

Fig. 1: Overview of DyFusion. The inference receives data from sensors and outputs probabilities over vessel labels. The fusion updates the sensor probabilities using contextual data.

Overview. DyFusion is depicted in Fig. 1 and consists of two steps: *inference* and *fusion*. In the inference, the inputs are RGB and IR images. We use RGB transformations to simulate additional sensors and to improve the classification accuracy. The data from each sensor is presented to a pre-trained *convolutional neural network* (CNN) with a logistic regression at the last layer. The inference outputs probabilities over maritime vessel classes for each input sensor.

In the fusion, we calculate contextual data from the inference (CNN activation norm, contrast, size and symmetry of the input images). The contextual data is used on a probabilistic model to determine a confidence score for the probabilities calculated in the inference. Lastly, confidence scores and inference probabilities are combined and used to update the probabilistic model of each input sensor in the fusion.

Contributions. The highlights of this work are:
1) Simulating additional sensor data using image transformations to improve the classification robustness.
2) Using contextual information to calculate classification confidence of the CNNs predictions.
3) Updating the confidence on CNNs based on the available contextual information.
4) Carrying out experiments to demonstrate the robustness of DyFusion when the test conditions change (image size, contrast, signal to noise ratio).

[1] For simplicity, we shortened the term to *sensor fusion* in this paper.
[2] Abbreviated to IR in the rest of this paper.

2. RELATED WORKS

Sensor fusion systems can be split considering whether fusion happens at the acquisition level [5, 6, 7, 8], at the feature description level [9, 10, 11, 12] or at the decision level [13, 14, 15]. We consider fusion at the decision level since the primary goal of DyFusion is to be robust to changes during operation and it is an easy strategy to simply discard sensor decisions with low confidence scores. We also calculate contextual information at each step of the decision pipeline for each sensor to estimate the prediction confidence and to update the fusion model. We briefly describe works that are closer to DyFusion in the next paragraphs. A comprehensive review of general sensor fusion can be found in [2].

The probabilistic model we use is similar to Beyca et. al. [6] work on non-parametric sensor fusion. Similarly, our probabilistic model aims at estimating a distribution of a target hidden random variable. In our case, the target is the optimal classification confidence conditioned to contextual information. In [6], the targets are drifting and anomaly on ultra-precision machining (UPM) conditioned to the sensor data. The non-parametric approach used by Beyca et. al. is based on the Dirichlet process and our approach is based on an ensemble of Gaussian distributions. The number of components in our Gaussian ensemble is a DyFusion parameter and the Dirichlet process may be considered to implement DyFusion without having to specify the number of components.

Li et. al. [13] use an ensemble of CNNs and a Dempster-Shafer (DS) approach to combine predictions from each CNN. DyFusion differs from [13] on the use of a probabilistic fusion approach. DS may produce counterintuitive results when representing incomplete data and when updating rules [16]. We also consider updating the fusion model, which is not considered in [13]. Methods to update the rules in DS can be found in [15] and may be used for comparison.

It is worth to mention at least one work about fusion of infrared and visible images since our work is evaluated on this type of data. Most works on IR and visible image fusion in the literature focus on face recognition due to IR images being capable of solving many issues in traditional face recognition [1]. The deep perceptual mapping from Sarfraz and Stiefelhagen [9] consists of projecting IR and visible images onto a common discriminative latent space by using CNNs as the mapping function. An alternative approach is to map each input sensor image into individual features and later use a feature combination approach [10]. As demonstrated on a recent work [17], both approaches present similar results, hence, we consider individual sensors since it is easier to devise an algorithm to update the confidence scores for each input sensor data in this manner.

There are many works on IR/visible data fusion that we do not discuss here due to lack of space. For a recent review on IR/visible face recognition, we recommend the survey by Ghiass et. al. [1].

3. TECHNICAL APPROACH

The following paragraphs refer to the steps presented in Fig 1 and briefly described in Section 1. The following notation is used in this section:

y : class label (ground truth).
$\hat{y}(i)$: predicted label from the i-th element of $\vec{p}$.
S : set of input sensors.
$\vec{p}$: probability vector of labels from the fusion step.
$\vec{p^s}$: probability vector of labels from the s-th sensor.
$\vec{\alpha^s}$: confidence score of labels from the s-th sensor.
C^s : set of contextual values calculated for sensor s.

Image transformations. The idea is to simulate additional data for the sensors by applying simple image transformations. Our hypothesis is that the additional data will increase the classification accuracy by providing robustness to lighting conditions even though the transformed images are highly correlated. We consider only the visible images and the transformations are from RGB to grayscale, each of the visible components (red, green and blue) and the pixel intensity ratio between blue and green (B/G), blue and red (B/R), and green and blue (G/R). The intuition for using the ratio between visible components is due to its invariance to lighting conditions.

Convolutional Neural Network (CNN). We use a similar approach as the one described in [3]. The CNN architecture is the *very deep convolutional network* with 19 layers (VGG19) [18] pre-trained on the ImageNet dataset [19]. We also tested with the VGG16 and ResNET architectures but they resulted in lower accuracy. We ignore the last layer (softmax) by taking the output of the last convolutional (max pooling with $25,088$ features) as features to train a maximum margin logistic regression [20]. We normalize the features to unit length and we use the original norm of the feature as one of the contextual variables. The IR, RGB and transformed images are presented to VGG19 even though the network was not pre-trained on the IR or in the transformed images. Images are resized to 224 by 224 pixels regardless of the aspect ratio. The parameters used are summarized in Table 1.

Image size:	$(224, 224, 3)$
Class weight:	$1/m_i$ (m_i: samples from class i)
Cost parameter (C):	1024
Regression type:	L2-regularized logistic regression
VGG19 features:	$25,088$ – ReLU (range $[0, \infty]$)

Table 1: CNN and logistic regression parameters.

Contextual information. We calculate four contextual variables: image size, contrast, symmetry and norm of the activations from CNN. Image size is simply width $\times$ height and the intuition is that the image size provides information on whether there is enough information in the image to classify it. Contrast is calculated as the variance of pixel intensities in the grayscale version of the input image: $(|I|-1)^{-1}\sum_{i\in I}(i-\bar{i})^2$, where I is set of pixel intensities and $\bar{i}$ is average intensity of I. Symmetry is calculated by reflecting a sliding

window horizontally in the image and taking the sum of absolute differences between average intensities in each window (image size 224×224, window size 32×32, stride 8×8). The contextual values v are normalized such that they lie on the interval $[0, 1]$. We use a *min-max* normalization: $\max(0, \min(1, (v - v_{min})/(v_{max} - v_{min})))$, where v_{min} and v_{max} are calculated using the training samples. The p_i^s values are also modeled as if they were contextual values.

Sensor fusion model. The idea is to calculate α_i^s as $\Pr\big((\hat{y}(i) = y) \cap C^s\big)$ and to combine it with $\vec{p^s}$ to calculate $\vec{p}$. Specifically, $\vec{p} = \sum_{s \in S} \vec{p^s} \odot \vec{\alpha^s}$, where $\vec{p}$ sums to one and $\odot$ denotes element-wise multiplication. The intuition is that we should probably lower p_i^s if we observe a small chance of $\hat{y}(i) = y$ and a particular C^s. For instance, if we observe that the predicted label is often wrong if the contrast of the image is low, even if p_i^s is close to one, we should probably set α^s such that p_i^s is closer to 0. In practice, we observe a confusion in p^s among multiple labels and we use the contextual information to reduce the confidence of some predictions based on contextual information we observed during training and test. Given $\vec{p}$, the predicted label is calculated as $\hat{y}\big(\arg\max_i(\vec{p_i})\big)$. The confidence score is calculated as:

$$\alpha_i^s = \sum_{c \in C^s} \sum_{b=1:B} \delta_{i,b,s} \exp\Big(-\frac{(c - \mu_b)^2}{2\sigma^2}\Big), \qquad (1)$$

where $\delta_{i,b,s} = \Pr\big(\hat{y}(i) = y | c\big)$. Equation 1 is derived from $\Pr\big((\hat{y}(i) = y) \cap C^s\big)$. $\delta_{i,b,s}$ is estimated from the training set using cross-validation and it is updated for each test sample following the *sensor confidence update* step described in the next paragraph. We use an ensemble of B Gaussian distributions with fixed variance $\sigma^2 = 0.01$ and mean uniformly distributed in the interval $[0, 1]$ (assuming $0 \leq c \leq 1$). We evaluated $B \in (5, 10, 15, 30)$ and found no significant difference in the results with B between 5 and 15. Hence, we assign $B = 5$ in this work.

Sensor confidence update. The goal is to estimate $\delta_{i,b,s}$ in Equation 1. We initialize the probability $\delta_{i,b,s}$ as 0.5 and we use an exponentially weighted rule to update $\delta_{i,b,s}$ for training and test samples. Specifically, $\delta_{i,b,s}^{t+1} = \delta_{i,b,s}^{t} \lambda + (1 - \lambda) z$, where $\lambda \in \mathbb{R}, 0 \leq \lambda \leq 1$, is the forgetting factor and z is the observed probability to be updated. Updating values is carried slightly different between training and test. During training, λ is empirically set to 0.95 and z to 1, if the training sample matches the ground truth, or z to 0, otherwise. To avoid over-fitting, we use ten-fold cross validation to estimate $\vec{p^s}$ in the training. During test, we set λ to either 1 if we are experimenting with the fusion model fixed, i.e., using only the training samples, or $\lambda = 0.80$ otherwise. We evaluated lambda in $0.95, 0.90, 0.80, 0.70$ and we found 0.80 to provide the best results. In the test, z is set as:

$$z = \begin{cases} max(\vec{p_i}), & \text{if } \hat{y}\big(\arg\max_i(\vec{p_i})\big) = \hat{y}\big(\arg\max_i(\vec{p_i^s})\big), \\ 1 - max(\vec{p_i}), & \text{otherwise.} \end{cases}$$

The idea is to update $\delta_{i,b,s}$ using $z = \max(\vec{p_i})$ calculated by the fusion model if the sensor pipeline predicted correctly, and $z = 1 - \max(\vec{p_i})$, otherwise. In case the CNN pipeline from sensor s and the fusion predicted differently from each other, we also update $\delta_{i,b,s}$ with $\max(\vec{p_i})$ for i given by $\arg\max_i(\vec{p_i})$. Finally, after each update, we normalize $\delta_{i,b,s}$ such that the probability over all i given b and s sums to 1.

4. EXPERIMENTAL RESULTS

We evaluate DyFusion on the VAIS dataset [3], which consists of $2,865$ cropped maritime vessel images ($1,623$ RGB and $1,242$ IR). An IR/RGB pair of a sample from VAIS is illustrated in Fig. 1. The RGB and IR images are synchronously captured from a pair of fixed cameras on land. Vessel images are captured in close distance, occasionally with the vessel docked at a pier such that high-resolution images are present in the dataset. Vessels may also appear far in open-sea such that there are small blurred images in the dataset. The number of pixels ranges from 644 to 4.47 million for RGB images and from 594 to 0.13 million for IR images.

The images are captured at various times of day, including dusk and dawn, such that some RGB images may appear dim and hard to recognize even with manual inspection. The automatic identification system (AIS) from nearby vessels and manual inspection are used to annotate the VAIS dataset. The dataset is split into 539 RGB/IR pairs and 334 singletons for training and 549 pairs and 358 singletons for the test. We use the average of accuracy per vessel label as the evaluation measurement following the VAIS protocol.

We compare our approach with three baselines: summing p^s from RGB and IR (called RGB+IR), convex combination of RGB and IR with 0.8 weight to RGB (0.8RGB+0.2IR), and the convex combination of RGB, transformations (T) and IR (called 0.06(RGB+T)+0.5IR). Weights are computed by maximizing the test accuracy as the upper-bound of the baseline. We also evaluated DyFusion using only training samples to compare with other works in the literature ($\lambda = 1$).

The following is the process we use to assess the dynamic aspect of DyFusion. We initialize the probabilistic model with the validation data according to Section 3. Then, we set $\lambda = 0.8$ and run DyFusion once for each test sample. We modify the test samples by reducing the scale, the contrast or the signal to noise ratio (SNR) of either RGB or IR (only one sensor each time). The modifications change the scale, contrast, and noise (in this order, one variable each time) from their original values to the worst-case scenario and then back to their original value. The goal is to evaluate if DyFusion can recover from the worst-case scenario and if the performance does not degrade over time.

In the dynamic experiments, the test images are scaled down to $1.00, 0.75$ and 0.50 of the input size. Contrast changes by applying gamma normalization to the image with gamma in $(1, 2.5, 5, 10, 15, 20, 25)$. SNR changes by summing random values with zero mean and standard deviation in $(0, 1, 5, 10, 15)$ to the pixel intensities. We repeat each experiment 30 times and we report the 0.95% confidence interval.

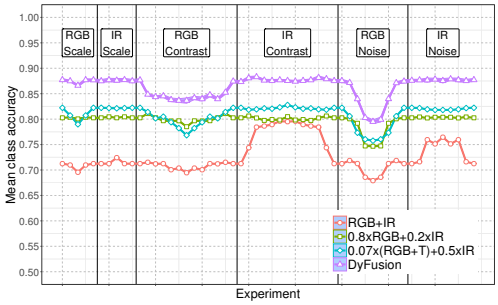

Fig. 2: DyFusion and baseline accuracy in different experiments.

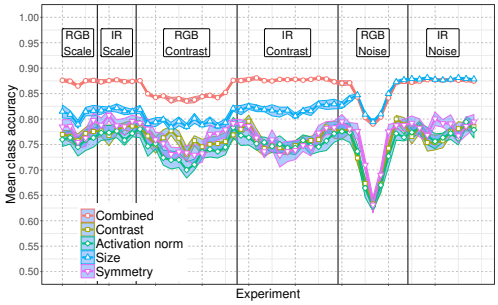

Fig. 3: DyFusion accuracy per contextual variable.

Baseline comparison. Results from DyFusion and the baseline approaches are shown in Fig 2. The simple approach of adding RGB and IR probabilities have the worst performance. Introducing weights to RGB and IR probabilities in 0.8xRGB+0.2xIR increases the accuracy by about 10 percentage points (p.p.). Adding the simulated sensor data in (0.06x(RGB+T)+0.5xIR) slightly increases the accuracy (about 2 p.p.) with the accuracy dropping lower than 0.8xRGB+0.2xIR in two situations: when the RGB is scaled down and when the contrast decreases. Adding contextual information (DyFusion) addresses most of the issues with (0.06x(RGB+T)+0.5xIR) by increasing the mean class accuracy by about 5 p.p. Notably, the accuracy of DyFusion never drops below any other approach. DyFusion is also surprisingly stable and it recovers the initial recognition performance after each worst-case scenario.

Context information evaluation. Fig. 3 decomposes the performance of DyFusion for each contextual variable. We observe more variance in the results for individual context variables compared to the combination. The symmetry, contrast and activation norm present similar performances. The image size provides the best individual performance and, interestingly, it converges to the same performance as the combination after some iterations. Upon inspection, we observe some correlation between image size and a few vessel labels (docked cargo ships have larger images) which can explain this result. Combining all four contextual variables provides the best accuracy for almost all scenarios.

Random samples. Instead of providing modified samples in sequence to DyFusion, we evaluate the scenario where random modifications are applied to either the RGB or the IR image (only one sensor each time, selected randomly). The

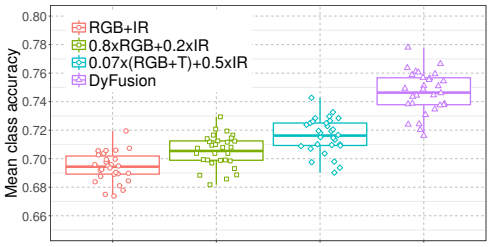

Fig. 4: Mean class accuracy for 54 experiments where random scale, SNR and contrast values are applied to test samples.

Method	Mean class accuracy
DyFusion ($\lambda = 0.8$)	0.882 ± 0.002
DyFusion ($\lambda = 1.0$)	0.873 ± 0.002
CNN+Gnostic Fields [3]	0.874
MFL (feature-level) + ELM [21]	0.876

Table 2: Comparison with other results reported on the VAIS dataset.

random modifications can be any combination of scale, SNR, and contrast values among the values presented in the 5th paragraph of Section 4. We run the experiment 54 times on all test samples without re-training DyFusion. Each time, we pick a different scale, contrast and noise parameter for the test samples. Results of the 54 experiments are in Fig. 4 and we conclude that, even in a completely random scenario, DyFusion is able to present a better performance on average.

Literature comparison. DyFusion with $\lambda = 1$ presents accuracy similar to the ones reported for the VAIS dataset (Table 2). Updating the fusion model on each test sample with $\lambda = 0.8$ results in a consistent slightly better result (about 1 p.p., difference of 7 samples classified correctly). The "medium-other" is the worst performing class with 0.62 accuracy[3] which can be explained by the mixture of different types of vessels under the same label. Most misses are also with the pier as background which may be confusing the CNN.

5. CONCLUSIONS

We investigated simulating sensors using image transformations which is shown to improve the results. To avail of the additional data from the multiple sensors, we proposed a fusion approach that estimates the probability that a prediction from a sensor is correct based on contextual information. By using both simulated sensor data and contextual information, we showed improved results compared to the baseline approaches. Finally, we demonstrated that the fusion approach can be updated on-the-fly using test samples and that it can recover from "hard" scenarios. In future works, we plan to investigate DyFusion's performance in other tasks and with different sensors.

6. ACKNOWLEDGMENT

This work was supported in part by NSF grant 1330110 and ONR grant N00014-12-1-1026. The contents of the information do not reflect the position or policy of US Government.

[3] An interactive confusion matrix is available at https://goo.gl/vD2QSf.

7. REFERENCES

[1] Reza Shoja Ghiass, Ognjen Arandjelović, Abdelhakim Bendada, and Xavier Maldague, "Infrared face recognition: A comprehensive review of methodologies and databases," *Pattern Recognition*, vol. 47, no. 9, pp. 2807–2824, 2014.

[2] Raffaele Gravina, Parastoo Alinia, Hassan Ghasemzadeh, and Giancarlo Fortino, "Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges," *Information Fusion*, vol. 35, pp. 68–80, 2017.

[3] Mabel M Zhang, Jean Choi, Kostas Daniilidis, Michael T Wolf, and Christopher Kanan, "Vais: A dataset for recognizing maritime imagery in the visible and infrared spectrums," in *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 10–16.

[4] Martin Liggins II, David Hall, and James Llinas, *Handbook of multisensor data fusion: theory and practice*, CRC press, 2017.

[5] Jiayi Ma, Chen Chen, Chang Li, and Jun Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Information Fusion*, vol. 31, pp. 100–109, 2016.

[6] Omer F Beyca, Prahalad K Rao, Zhenyu Kong, Satish TS Bukkapatnam, and Ranga Komanduri, "Heterogeneous sensor data fusion approach for real-time monitoring in ultraprecision machining (upm) process using non-parametric bayesian clustering and evidence theory," *Transactions on Automation Science and Engineering*, vol. 13, no. 2, pp. 1033–1044, 2016.

[7] Qi Wei, Nicolas Dobigeon, and Jean-Yves Tourneret, "Bayesian fusion of multi-band images," *Journal of Selected Topics in Signal Processing*, vol. 9, no. 6, pp. 1117–1127, 2015.

[8] Hongguang Li, Wenrui Ding, Xianbin Cao, and Chunlei Liu, "Image registration and fusion of visible and infrared integrated camera for medium-altitude unmanned aerial vehicle remote sensing," *Remote Sensing*, vol. 9, no. 5, pp. 441, 2017.

[9] M Saquib Sarfraz and Rainer Stiefelhagen, "Deep perceptual mapping for thermal to visible face recognition," *arXiv preprint arXiv:1507.02879*, 2015.

[10] Benjamin S Riggan, Nathaniel J Short, and Shuowen Hu, "Optimal feature learning and discriminative framework for polarimetric thermal to visible face recognition," in *Applications of Computer Vision (WACV), Winter Conference on*. IEEE, 2016, pp. 1–7.

[11] Shuowen Hu, Jonghyun Choi, Alex L Chan, and William Robson Schwartz, "Thermal-to-visible face recognition using partial least squares," *JOSA A*, vol. 32, no. 3, pp. 431–442, 2015.

[12] Zhongli Ma, Jie Wen, Quanyong Liu, and Guanjun Tuo, "Near-infrared and visible light image fusion algorithm for face recognition," *Journal of Modern Optics*, vol. 62, no. 9, pp. 745–753, 2015.

[13] Shaobo Li, Guokai Liu, Xianghong Tang, Jianguang Lu, and Jianjun Hu, "An ensemble deep convolutional neural network model with improved ds evidence fusion for bearing fault diagnosis," *Sensors*, vol. 17, no. 8, pp. 1729, 2017.

[14] Fernando Alonso-Fernandez and Josef Bigun, "Near-infrared and visible-light periocular recognition with gabor features using frequency-adaptive automatic eye detection," *IET Biometrics*, vol. 4, no. 2, pp. 74–89, 2015.

[15] Xinnan Fan, Pengfei Shi, Jianjun Ni, and Min Li, "A thermal infrared and visible images fusion based approach for multitarget detection under complex environment," *Mathematical Problems in Engineering*, vol. 2015, 2015.

[16] Judea Pearl, "Reasoning with belief functions: An analysis of compatibility," *Int. Journal of Approximate Reasoning*, vol. 4, no. 5-6, pp. 363–389, 1990.

[17] Shuowen Hu, Nathaniel J Short, Benjamin S Riggan, Christopher Gordon, Kristan P Gurton, Matthew Thielke, Prudhvi Gurram, and Alex L Chan, "A polarimetric thermal database for face recognition research," in *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 119–126.

[18] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255.

[20] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[21] Longhui Huang, Wei Li, Chen Chen, Fan Zhang, and Haitao Lang, "Multiple features learning for ship classification in optical imagery," *Multimedia Tools and Applications*, pp. 1–27, 2017.