DARE: AI-based Diver Action Recognition System using Multi-Channel CNNs for AUV Supervision

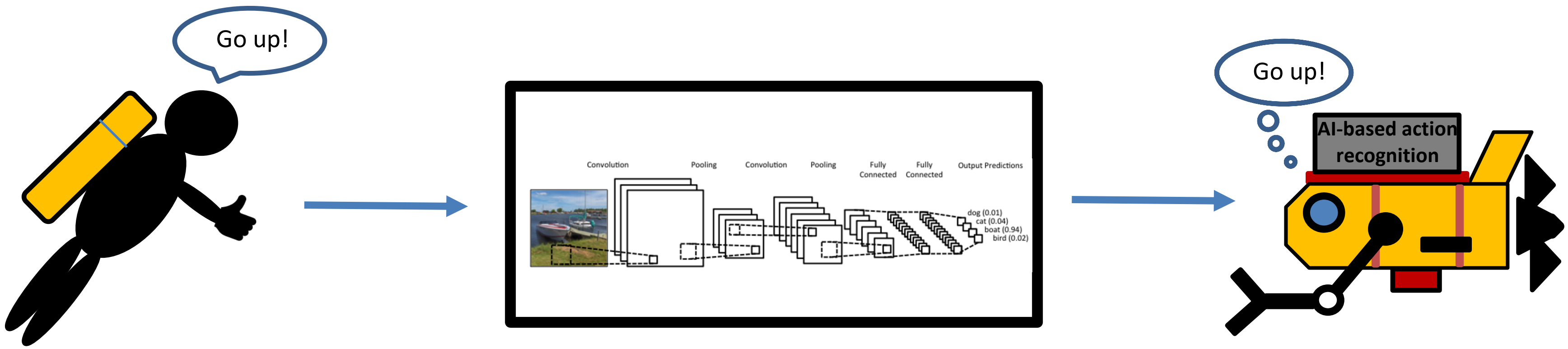
This work has been published in:

J. Yang, J. P. Wilson and S. Gupta. "DARE: AI-based Diver Action Recognition System using Multi-Channel CNNs for AUV Supervision", arXiv: 2011.07713, 2020

Diver Action Recognition for Underwater Human-Robot Interaction

Human Robot Interaction requires the ability to dynamically reprogram the robot's mission parameters and human control input is limited to visual command.

- ❑ Traditional commands require waterproof joystick, keyboard or tablet.
- ❑ Diver gestures/ pose commands are more convenient and faster.



Diver action signal | Deep learning model | AUV receive signal

Objective: Develop a deep learning model on diver's action images to perform action recognition faster and accurately

- ❑ Diver images collected in both open sea and swimming pool should be included.
- ❑ Minimize the classifying time for each action.
- ❑ High accuracy is required when classifying all the action types.

Literature Review

Existing Image Classification for Diver Gesture

"Dynamic reconfiguration of mission parameters in underwater human-robot collaboration," Islam 2018 [1]

- ❑ Images captured in ideal swimming pool and terrestrial environment
- ❑ RGB camera
- ❑ Convolution Neural network

"Gesture-recognition as basis for a human-robot interface (HRI) on an AUV," Buelow 2011[2]

- ❑ Images were captured in ideal swimming pool environment
- ❑ Monocular camera
- ❑ Motion trajectories

"Understanding human motion and gestures for underwater human-robot collaboration," Islam 2018[3]

- ❑ Images were captured in ideal swimming pool and open sea environment
- ❑ Monocular RGB camera
- ❑ Fast Recurrent Convolution Neural Network

"Underwater Motion and Activity Recognition using Acoustic Wireless Networks," Hu 2020[4]

- ❑ Simulated target body velocity using acoustic wireless networks
- ❑ Arm motions classification
- ❑ Convolution Neural network

Research Gap: None or Limited amount of images captured in real open sea environments. Images captured using monocular RGB camera contain bland spot (information lost). Underwater diver motion classification involved not only arm but also whole body.

[1] M. J. Islam, M. Fulton, and J. Sattar, "Dynamic reconfiguration of mission parameters in underwater human-robot collaboration," in Robotics and Automation Letters IEEE, vol. 4, (Brisbane, QLD, Australia), pp. 113–120, 2018.

[2] H. Buelow and A. Birk, "Gesture-recognition as basis for a human robot interface (HRI) on an AUV," in OCEANS'11 MTS/IEEE KONA, (Waikoloa, HI, USA), pp. 1–9, 2011.

[3] M. J. Islam, M. Ho, and J. Sattar, "Understanding human motion and gestures for underwater human-robot collaboration," J. Field Robot vol. 35, pp. 1–23, 2018.

[4] H. Hu, Z. Sun, and L. Su, "Underwater motion and activity recognitionusing acoustic wireless networks," inICC2020-2020IEEEInternationalConferenceonCommunications(ICC)., (Dublin, Ireland), 2020.

CADDY Dataset

Gestures

Cognitive autonomous diving buddy (CADDY) gesture which include open sea and swimming pool scenario with 16 different gestures and 3 poses to recognize.
Underwater diver postures are focus on whole body including arm positions.

- Data collected location: **open seas** of Biograd na Moru, Croatia, an **indoor pool** in the Brodarski Institute, Croatia, and an **outdoor pool** in Genova, Italy.
- 16 different gestures + 1 true negative
- Up, down, backwards, carry, boat, 1-4, take a photo etc
- Stereo pairs of gestures (9239+7190)*2=32,858

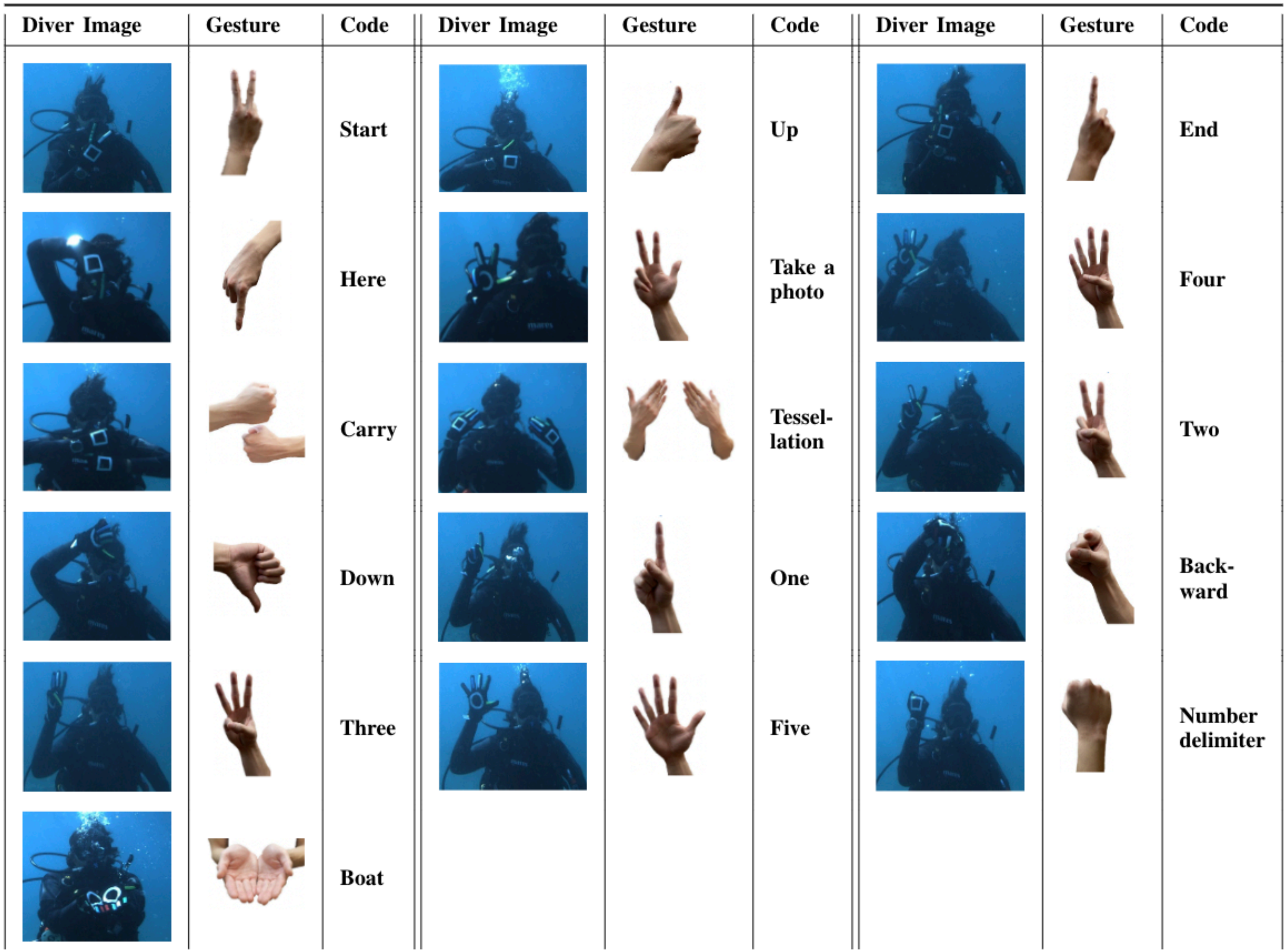
  9239→gestures, 7190→true negatives

Underwater diver gesture sample images in various environments

[5]. A. G. Chavez, A. Ranieri, D. Chiarella, E. Zereik, A. Babi, and A. Birk, "CADDY underwater stereo-vision dataset for human-robot interaction (HRI) in the context of diver activities,"Journal of Marine Science and Engineering, vol. 7, pp. 16–29,2019.

CADDY Dataset

Gesture Table



Diver Image	Gesture	Code	Diver Image	Gesture	Code	Diver Image	Gesture	Code
		Start			Up			End
		Here			Take a photo			Four
		Carry			Tessel-lation			Two
		Down			One			Back-ward
		Three			Five			Number delimiter
		Boat						

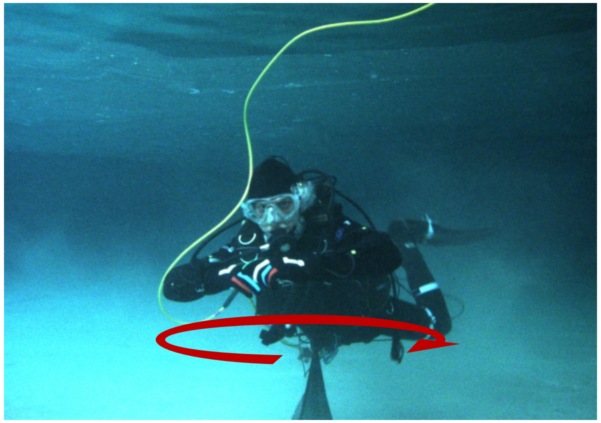
CADDY dataset diver hand gesture

CADDY Dataset

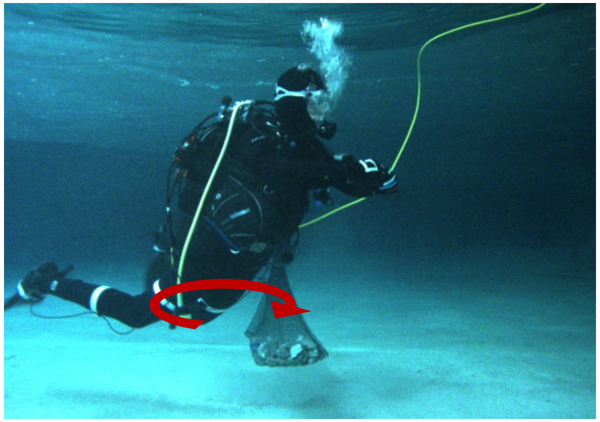
Poses

- Data collected location: **open seas** of Biograd na Moru, Croatia, an **indoor pool** in the Brodarski Institute, Croatia, and an **outdoor pool** in Genova, Italy.
- 3 different poses
- (1) turn horizontally (2) turning vertically (3) swim freely.

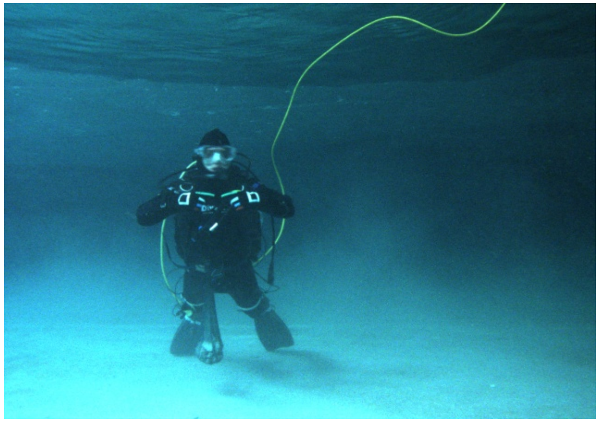
Stereo pairs of gestures (3934+2722+6052)*2=25,416



(1) Turning horizontally



(2) Turning vertically

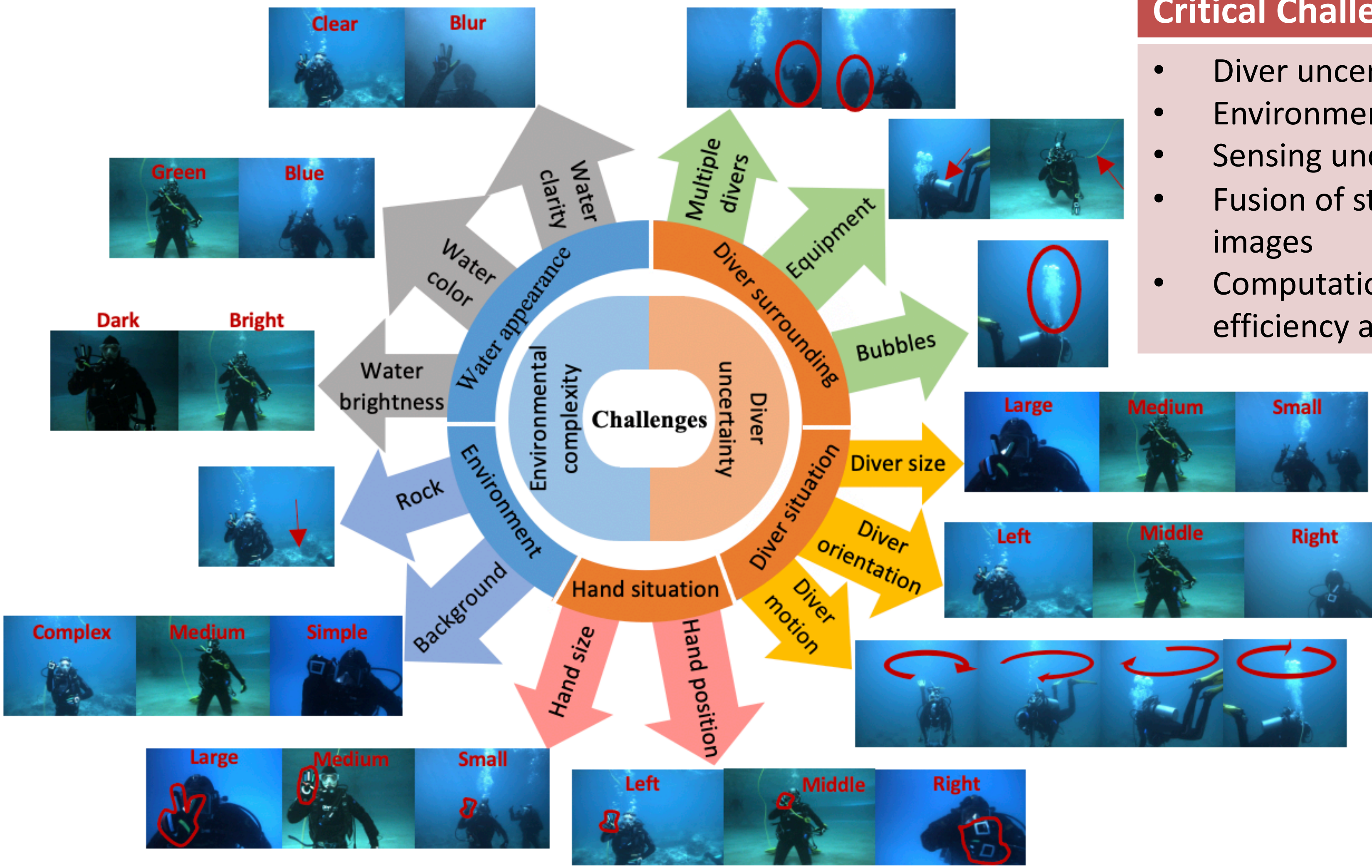


(3) swim freely

Underwater diver pose sample images

[5]. A. G. Chavez, A. Ranieri, D. Chiarella, E. Zereik, A. Babi, and A. Birk, "CADDY underwater stereo-vision dataset for human-robot interaction (HRI) in the context of diver activities,"Journal of Marine Science and Engineering, vol. 7, pp. 16–29,2019.

Challenges of Underwater Diver Action Recognition

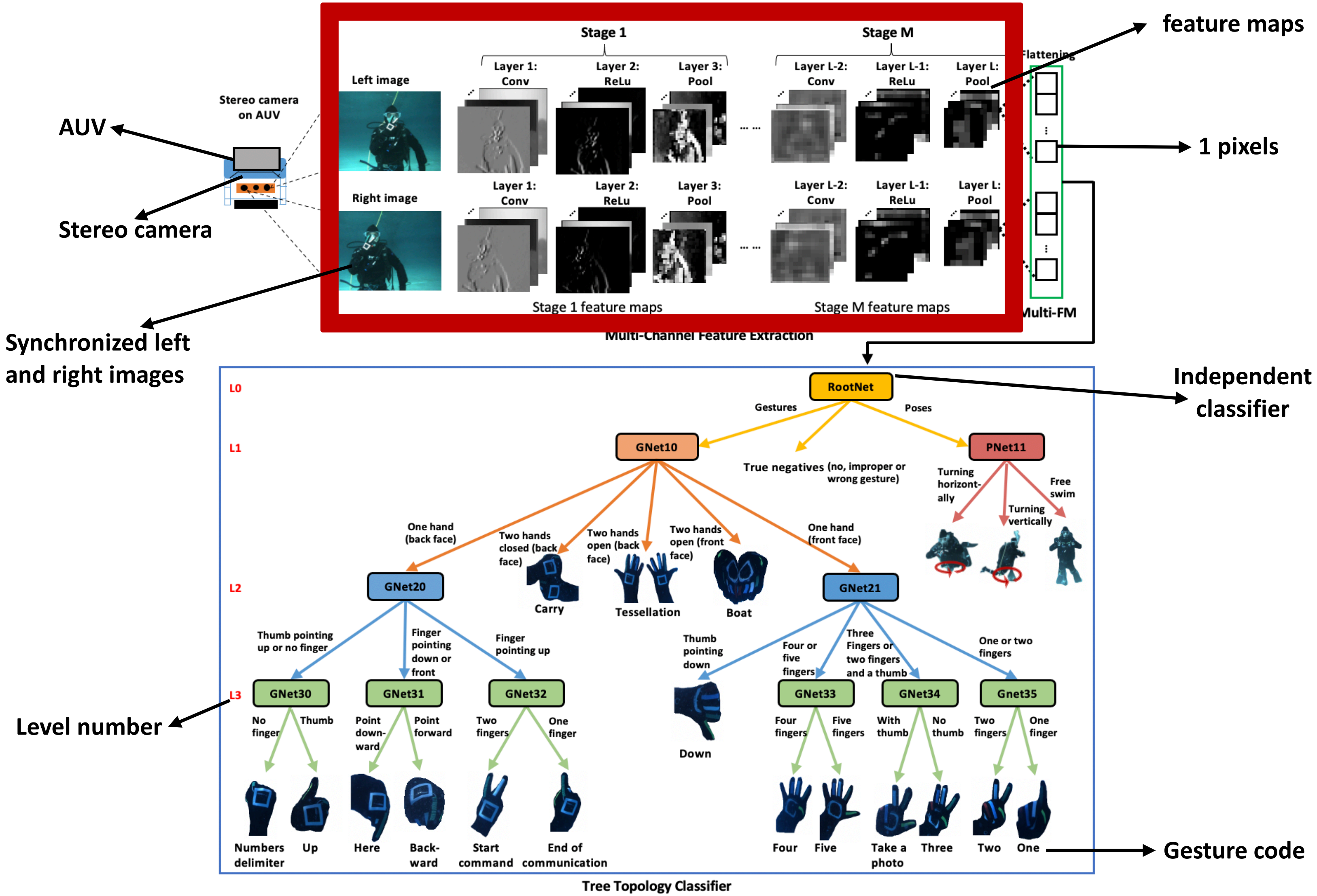


Challenges of diver uncertainty and environment complexity

Critical Challenges

- Diver uncertainty
- Environment complexity
- Sensing uncertainties
- Fusion of stereo camera images
- Computational efficiency and reliability

DARE Architecture



Training architecture

Transfer learning Feature Extraction Advantages

Main Idea: Use transfer learning pre-trained Convolutional neural network (CNN) to extract useful features for training and classifying the diver's gestures and poses.

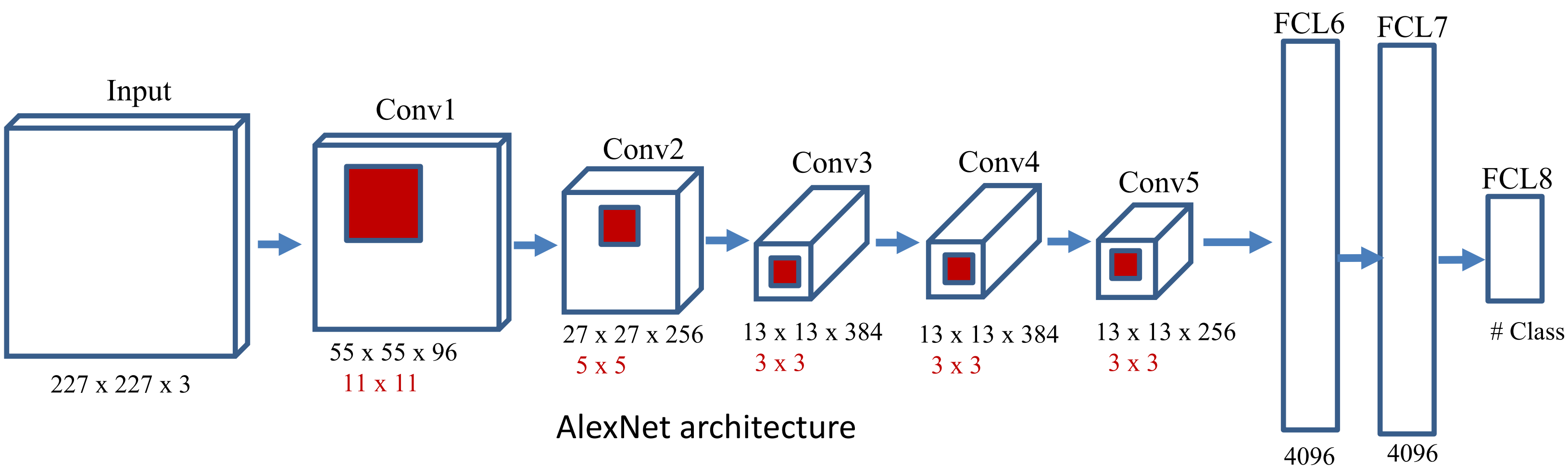
Model	Training Data	Computation	Training Time	Model Accuracy
Traditional CNN	1000s to millions of label images	Compute intensive	Days to weeks for realproblems	High (can overfit to small dataset)
Transfer learning pre-trained CNN	100s to 1000s of label images	Moderate computation	Minutes to hours	Good, depends on model structure

Benefits:

- ❑ Filters in convolution layer in CNN produce feature maps which contain important information. And Pooling layer will preserve the useful information but reduce the image size.
- ❑ Transfer learning prevents overfitting from training a network from scratch.
- ❑ Using different transfer learning nets provides a scope to observe the differences between the network structure and result.

Pre-Trained Network Architectures

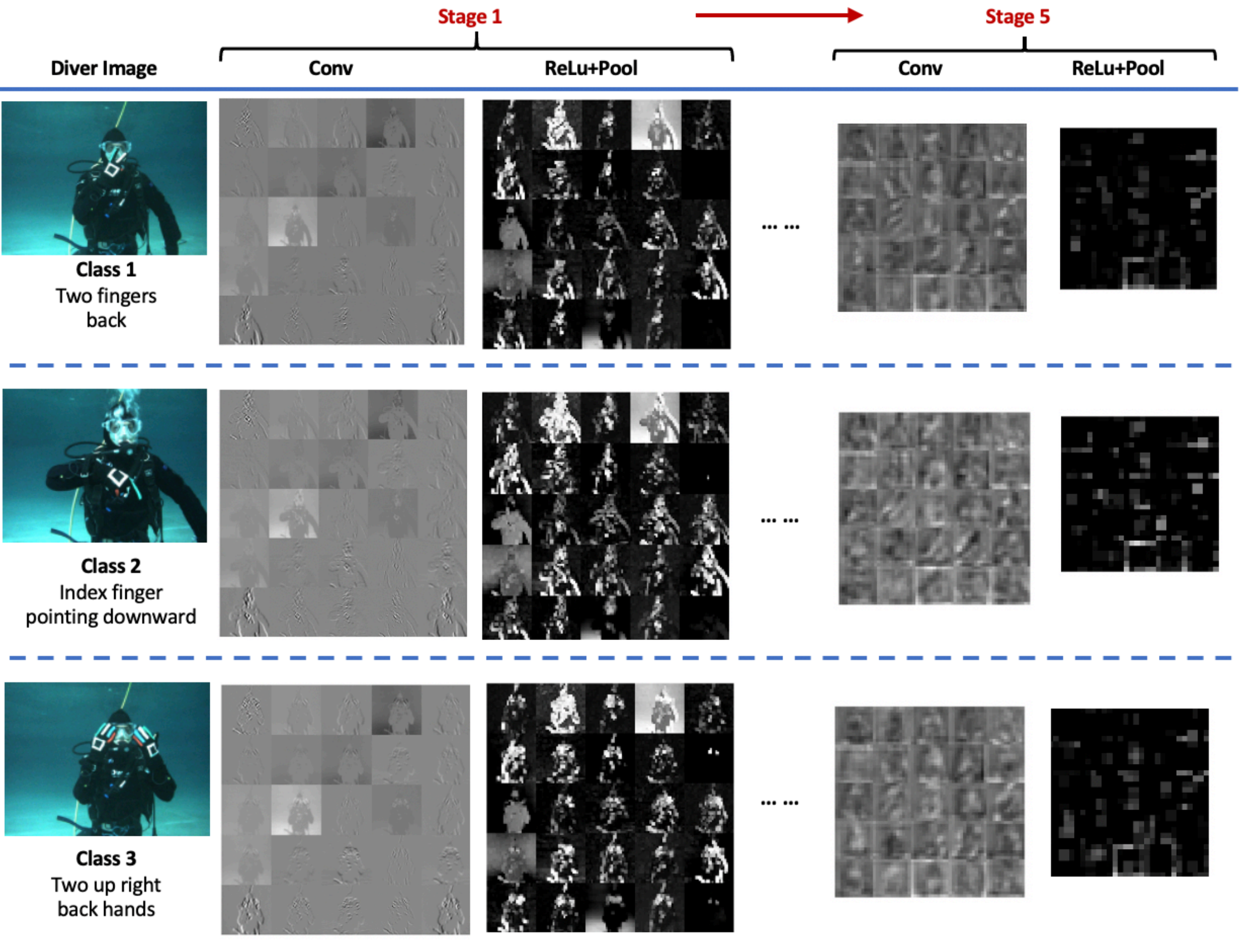
	AlexNet	VggNet	ResNet
Depth	8	16	18
Input Layer size	227x227x3	224x224x3	224x224x3
Filter Size	11x11, 5x5, 3x3	3x3	7x7, 3x3, 1x1
Number of Conv layer	5	13	17
Number of Fully-connected layer	3	3	1
Number of hyper-parameters	61.1 million	138.4 million	11.2 million



AlexNet architecture

[6]. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25 (NIPS), pp. 1097–1105, 2012.
[7]. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition,"arXiv:1409.1556,2014.
[8]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

Feature Visualization Using AlexNet

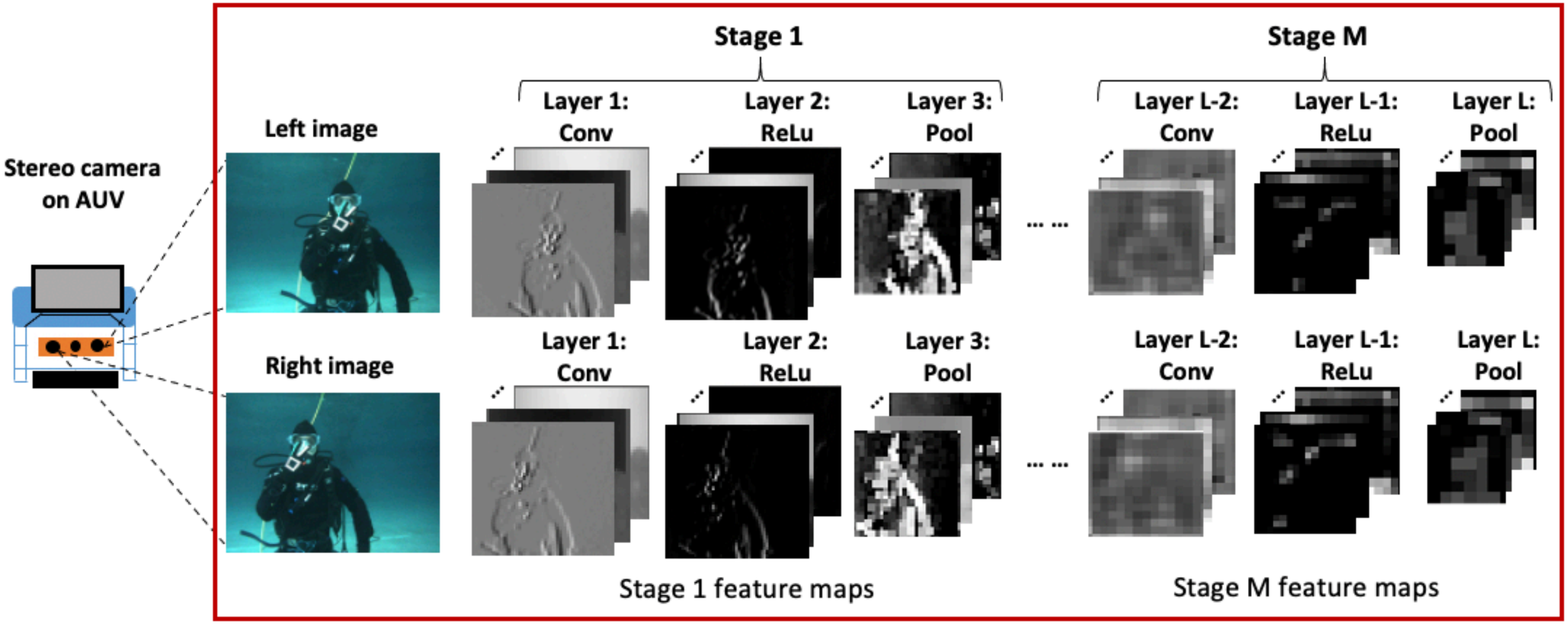


Features process:

- Stage 1 contains 96 features in total with selected 25 random feature maps shown in figure.
- **Convolutional layer** used filters to extract information across the images such as edge and line.
- **ReLu** is an activation function which introduce non-linearity to the network
- **Max Pooling layer** down-samples the images
- Stage 5 output feature maps contains difference among these three different gestures

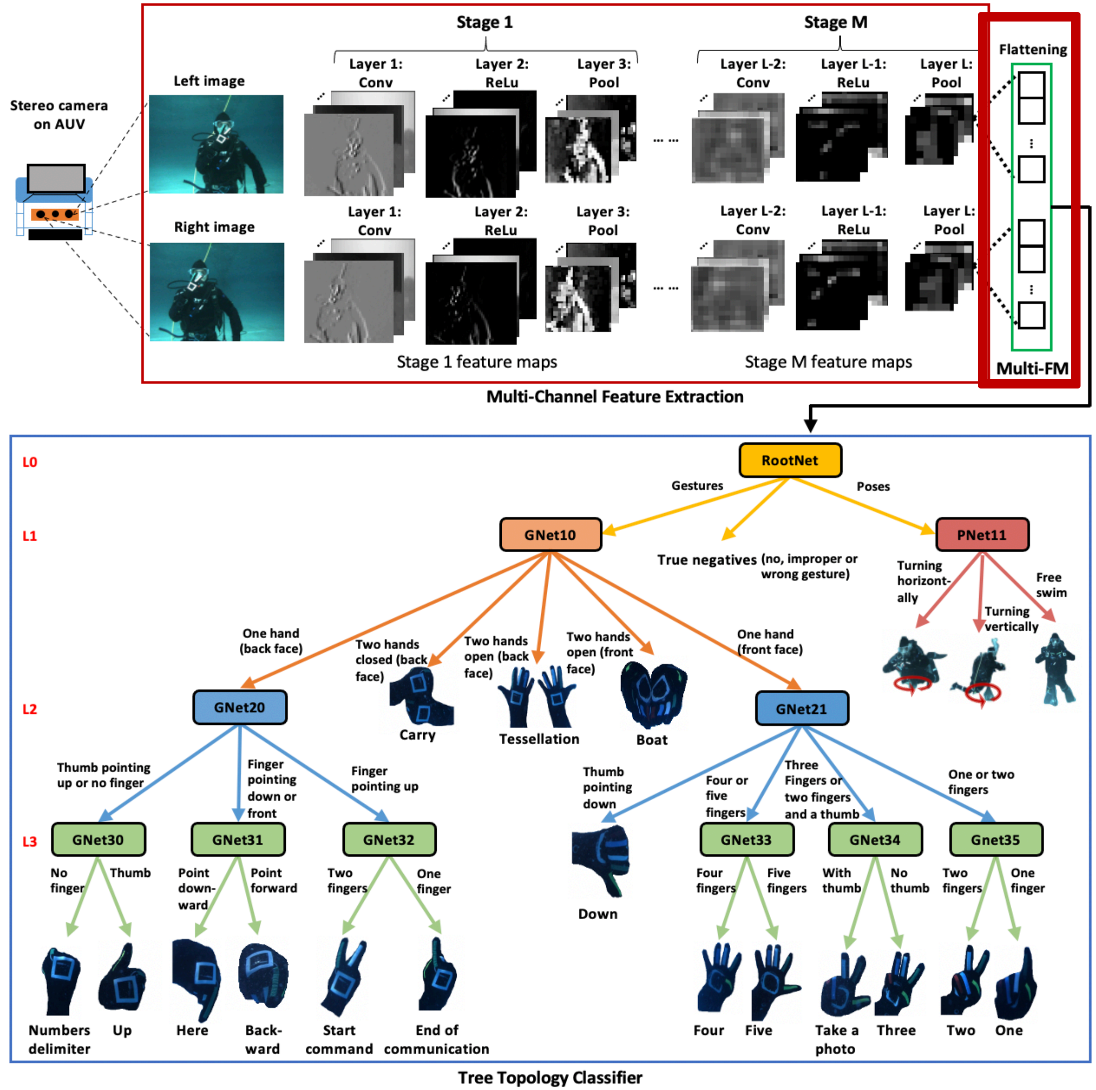
Objective

- Preserve correlation between left and right stereo images
- Maintain crucial diver AUV distance information
- Prevent overfitting



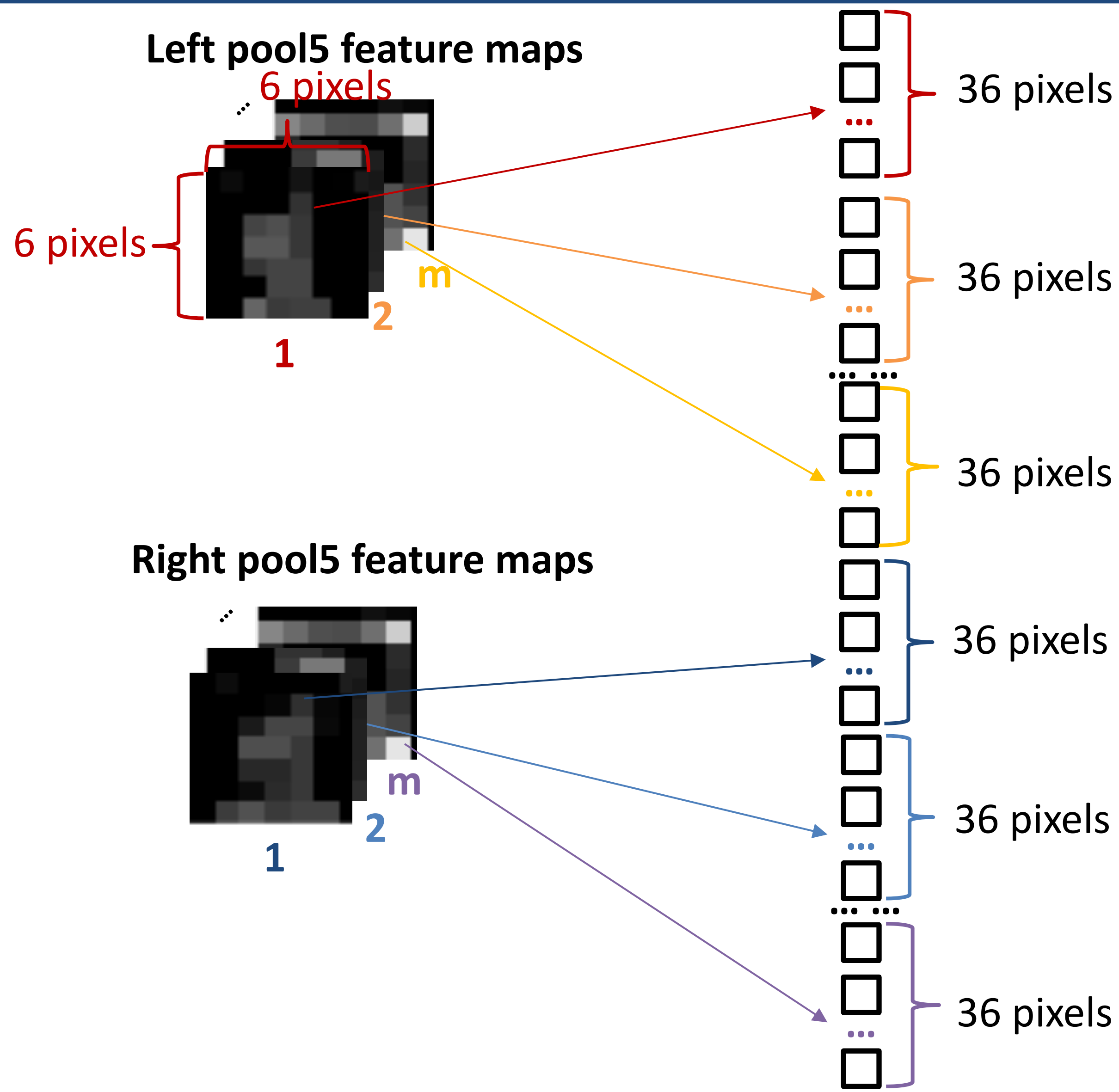
Multi-Channel Feature Extraction

DARE Architecture



Training architecture

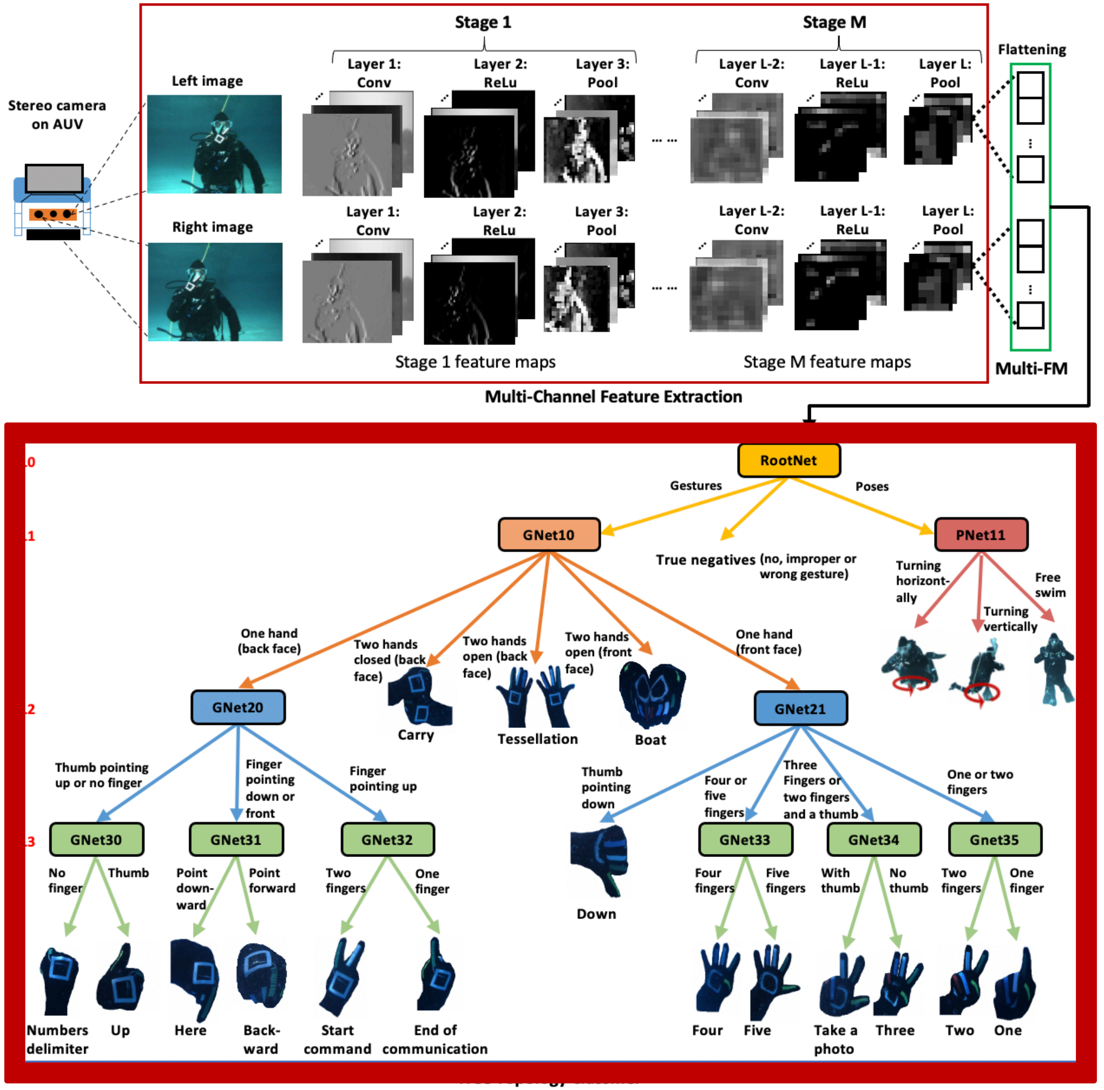
Flattening



Left flatten size: 6x6xm

Right flatten size: 6x6xm

DARE Architecture



Training architecture

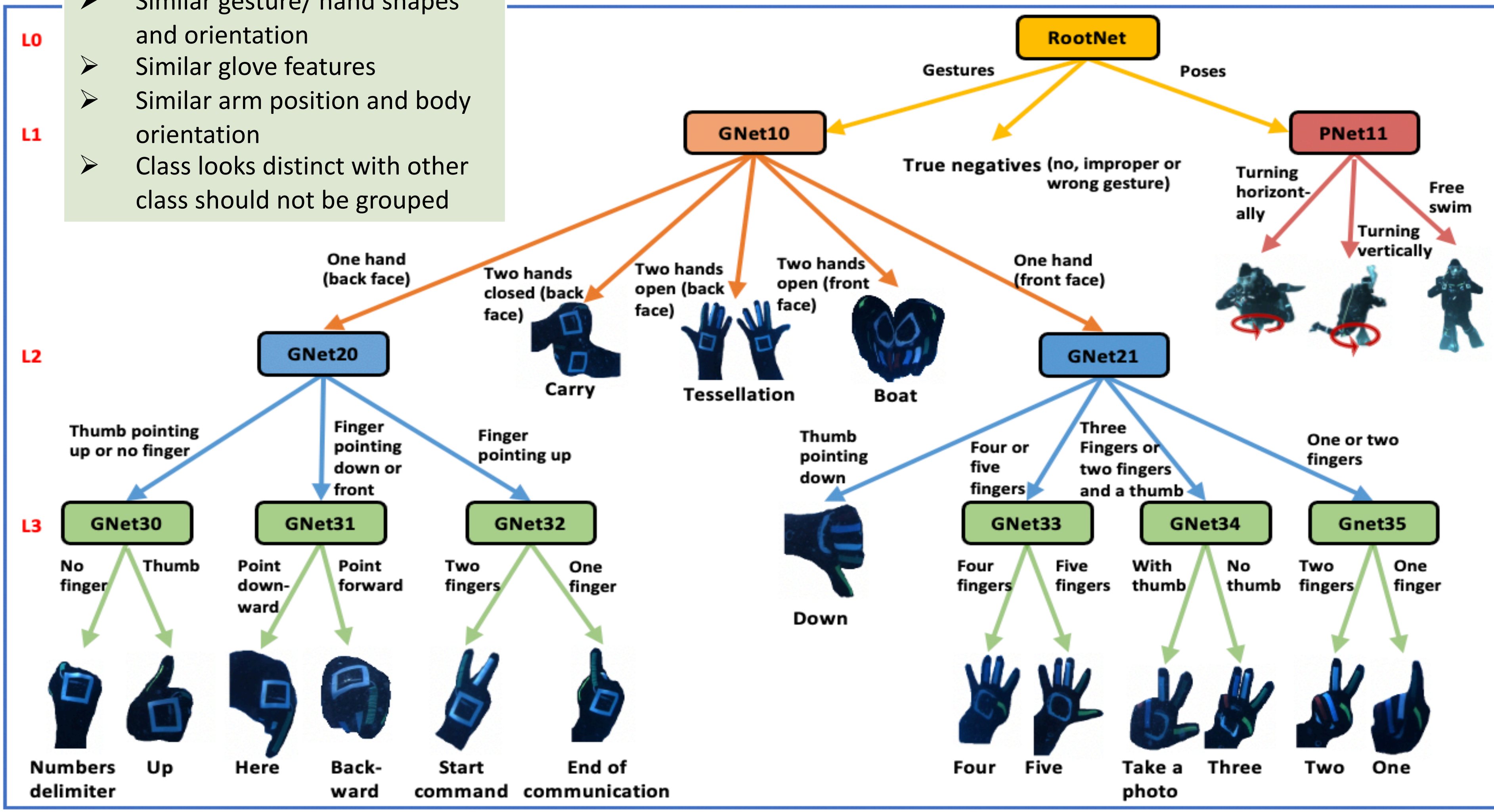
Hierarchy Tree-structured Neural Network Classifier

UCONN

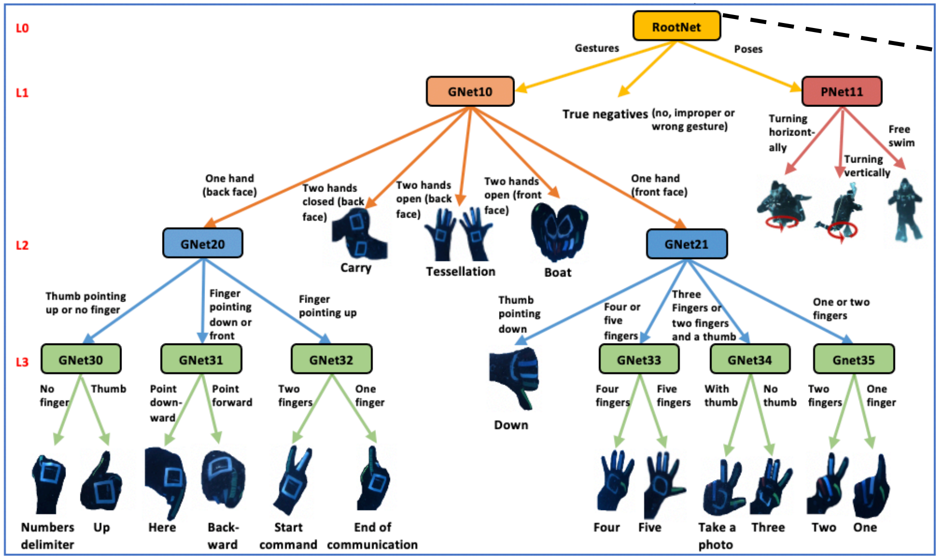
LINKS
Laboratory of Intelligent Networks and Knowledge-Perception Systems

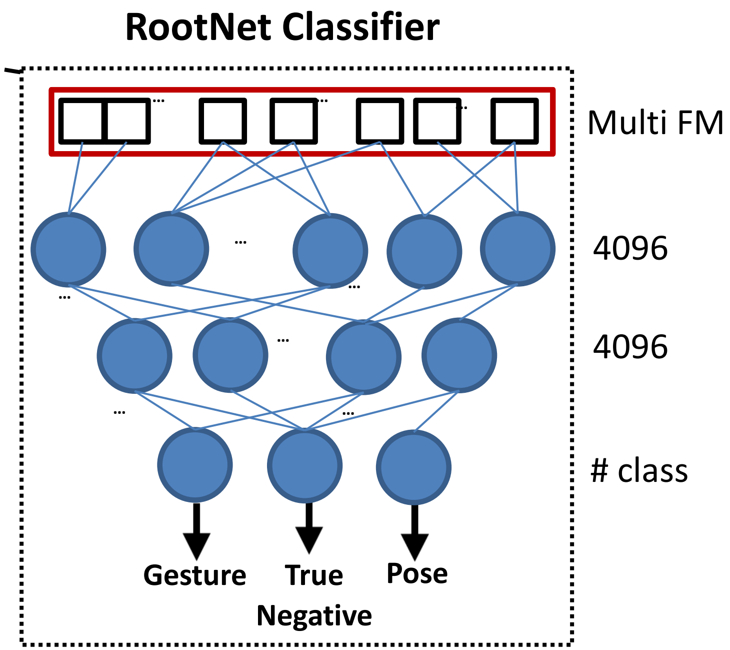
Grouping Standard

- ➢ Similar gesture/ hand shapes and orientation
- ➢ Similar glove features
- ➢ Similar arm position and body orientation
- ➢ Class looks distinct with other class should not be grouped



Hierarchy Tree-structured Neural Network Classifier





RootNet Classifier Architecture

Classifier Training Process

- Input: Multi FM with corresponding class, Output: Diver action class
- Train 11 small classifiers independently with same network architecture
- Parameters fine-tuning for each classifier

Classifier Testing Process

- Test feature extracted using multi-channel network first goes into RootNet
- Perform prediction on next level based on result from previous level
- Stop prediction when diver action class is the output

Analysis Setup

Prepressing procedure:

- 5-Fold cross validation with each fold 20% for testing and 80% for training
- Resize Image using down-sample ratio

Parameters:

- Solver: Stochastic gradient descent with momentum (SGDM)
- Initial learning rate:0.001
- minibatch size: 64 (default), 500/ 100 (RootNet, GNet10 and PNet11)

Machine:

- MATLAB Deep Learning Toolbox
- Windows 10 computer with an Intel Core i7 processor and 32GB RAM

Performance Measures

Confusion Matrix

Target Class \ Prediction Class	C0	C1	C2
C0	C00	C01	C02
C1	C10	C11	C12
C2	C20	C21	C22

$C_{i=0}$ confusion matrix

Target Class \ Prediction Class	C0	C1	C2
C0	TP	FN	FN
C1	FP	TN	TN
C2	FP	TN	TN

Correct Classification Rate:

$$TPR_i = \frac{TP_i}{TP_i + FNi} \qquad TNR_i = \frac{TN_i}{TN_i + FPi}$$

$$Balanced\ Individual\ \ Class\ Accuracyi = \frac{TPR_i + TNRi}{2}$$

$$Overall\ CCR = \frac{\sum_i C_{ii}}{\sum_i \sum_j C_{ij}}$$

F1 Score:

$$Precision_i = \frac{TP_i}{TP_i + FPi} \qquad Recall_i = TPR_i = \frac{TP_i}{TP_i + FNi}$$

$$F1\ Scorei = 2\times\frac{Precision_i \times Recalli}{Precision_i + Recalli}$$

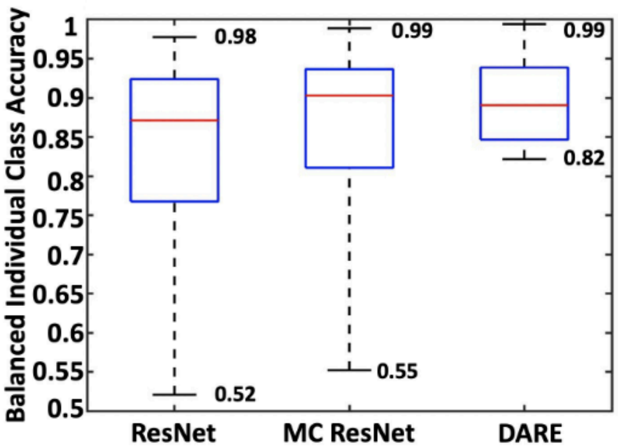
$$Overall\ F1\ Score = \frac{1}{N}\sum_{i=0}^{N} F1\ Scorei$$

Where N = total class number

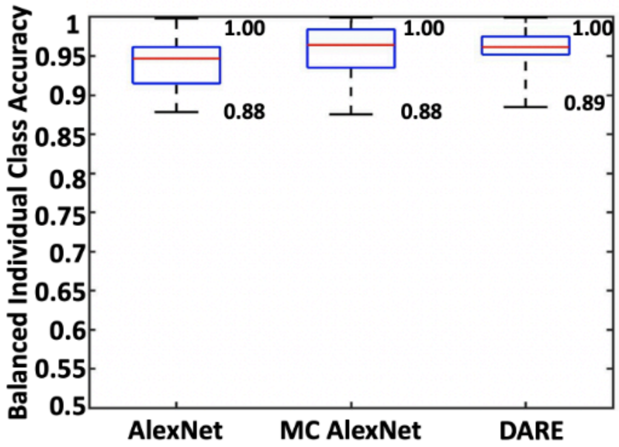
Results

Individual Class Performance

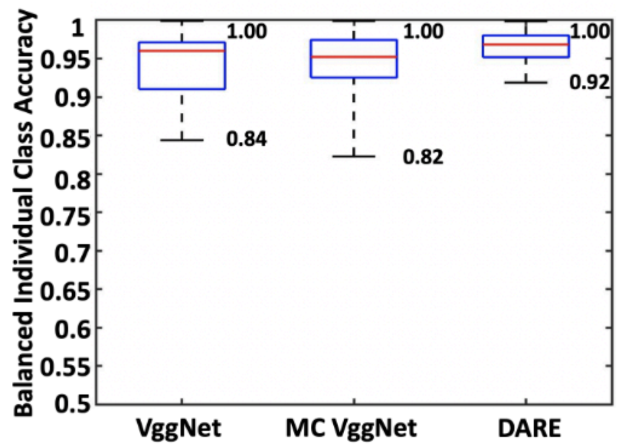
Balanced accuracies of ResNet-based networks



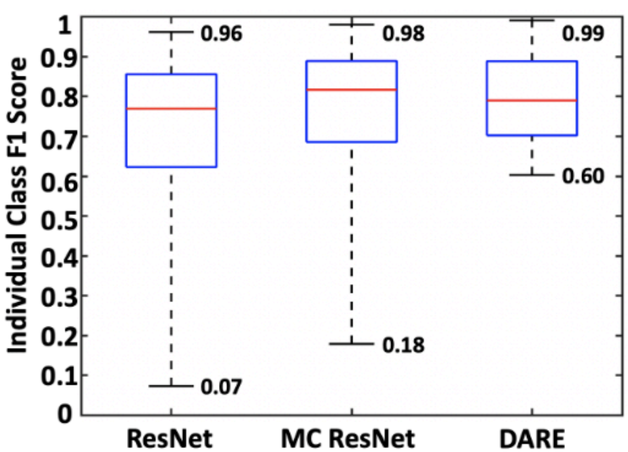
Balanced accuracies of AlexNet-based networks



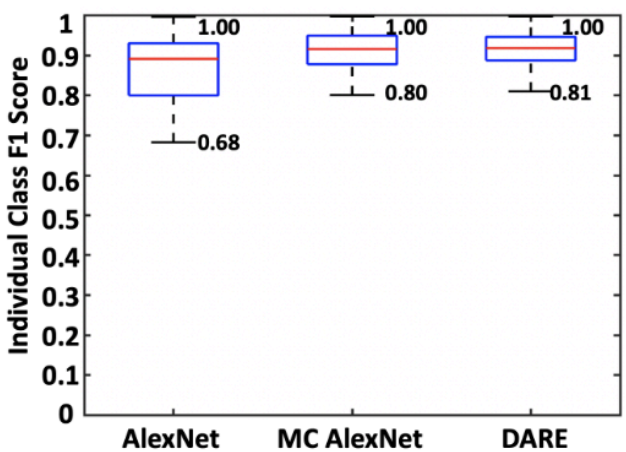
Balanced accuracies of VggNet-based networks



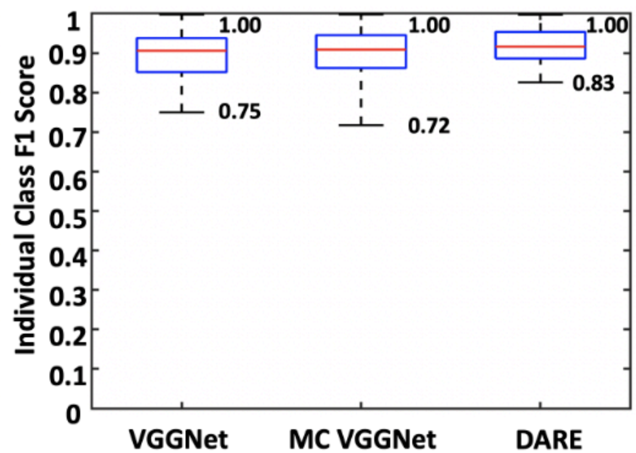
F1 scores of ResNet-based networks



F1 scores of AlexNet-based networks



F1 scores of VggNet-based networks



Results

Overall Performance

Metrics	ResNet			AlexNet			VggNet		
	Regular	MC	DARE	Regular	MC	DARE	Regular	MC	DARE
CCR (%)	86.03	88.80	89.47	94.21	95.88	95.93	95.20	95.39	95.87
F1 score	0.728	0.782	0.799	0.875	0.919	0.920	0.899	0.901	0.921
Training Time (hrs)	1.46	1.72	3.65	2.08	3.92	13.80	7.82	13.80	17.19
Testing Time (ms)	34.59	69.24	72.65	16.69	33.33	34.20	266.75	533.09	534.49

Conclusion and Future Work

Conclusion

- Human robot interaction application using DARE achieve high CCR equal to 95.87% in relatively short amount of time
- DARE boost every individual class accuracy above 92%
- Suitable for real-time application

Future Work

- Convolutional neural network grid search to find the best combination of hyper-parameters
- Image pre-processing techniques can be embedded to improve the algorithm performance
- Automate tree constructure for human-robot interaction