Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

# Comparative Evaluation of Adversarial Attacks and its Defense Mechanisms on Deep Neural Network

## R&D Defense

October 8, 2021


Prabhudev Bengaluru Kumar

*Advisors*
Prof. Dr. Nico Hochgeschwender,
M.Sc. Deebul Nair

# Table of Contents

# Table of Contents

# Introduction

*What is the adversarial attack?*

- **Adversarial attack** - injecting adversarial noise into input data to confuse a Deep Neural Network (DNN)

# Introduction

*What is the adversarial attack?*

- **Adversarial attack** - injecting adversarial noise into input data to confuse a Deep Neural Network (DNN)

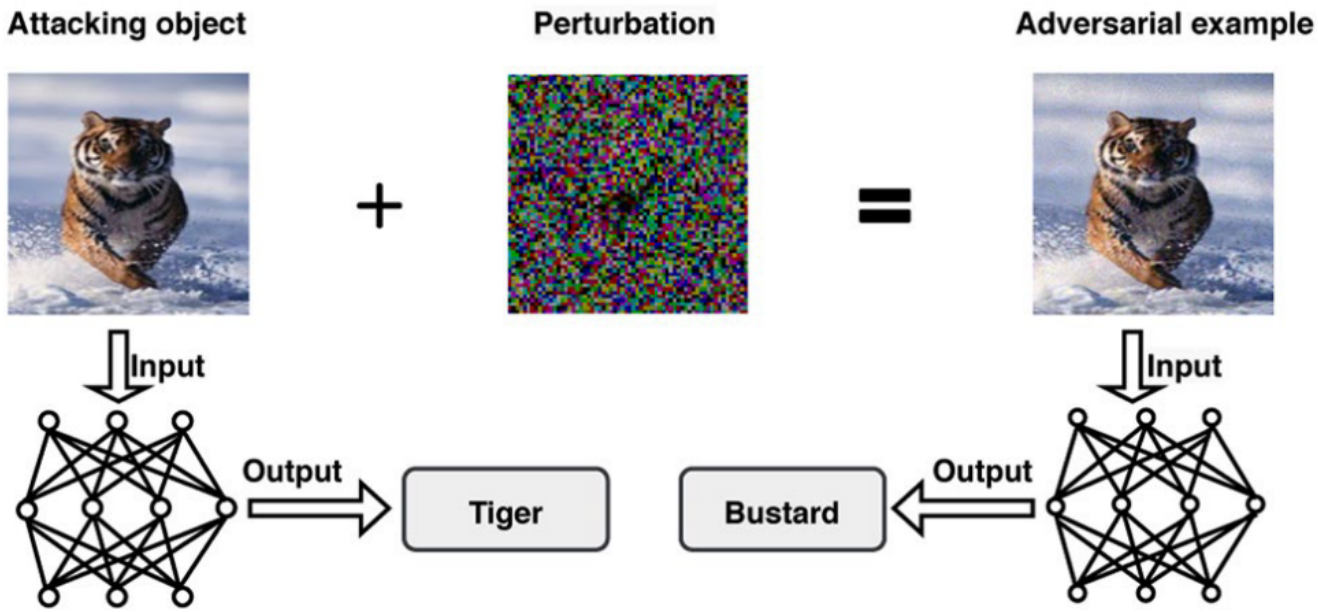



Figure 1: Example of adversarial attack. Image credits: [18]

# Introduction

*Different adversarial attack scenarios*

- **Digital world attack** - fast gradient sign method [8], Carlini and Wagner attacks [6], one pixel attack [21]

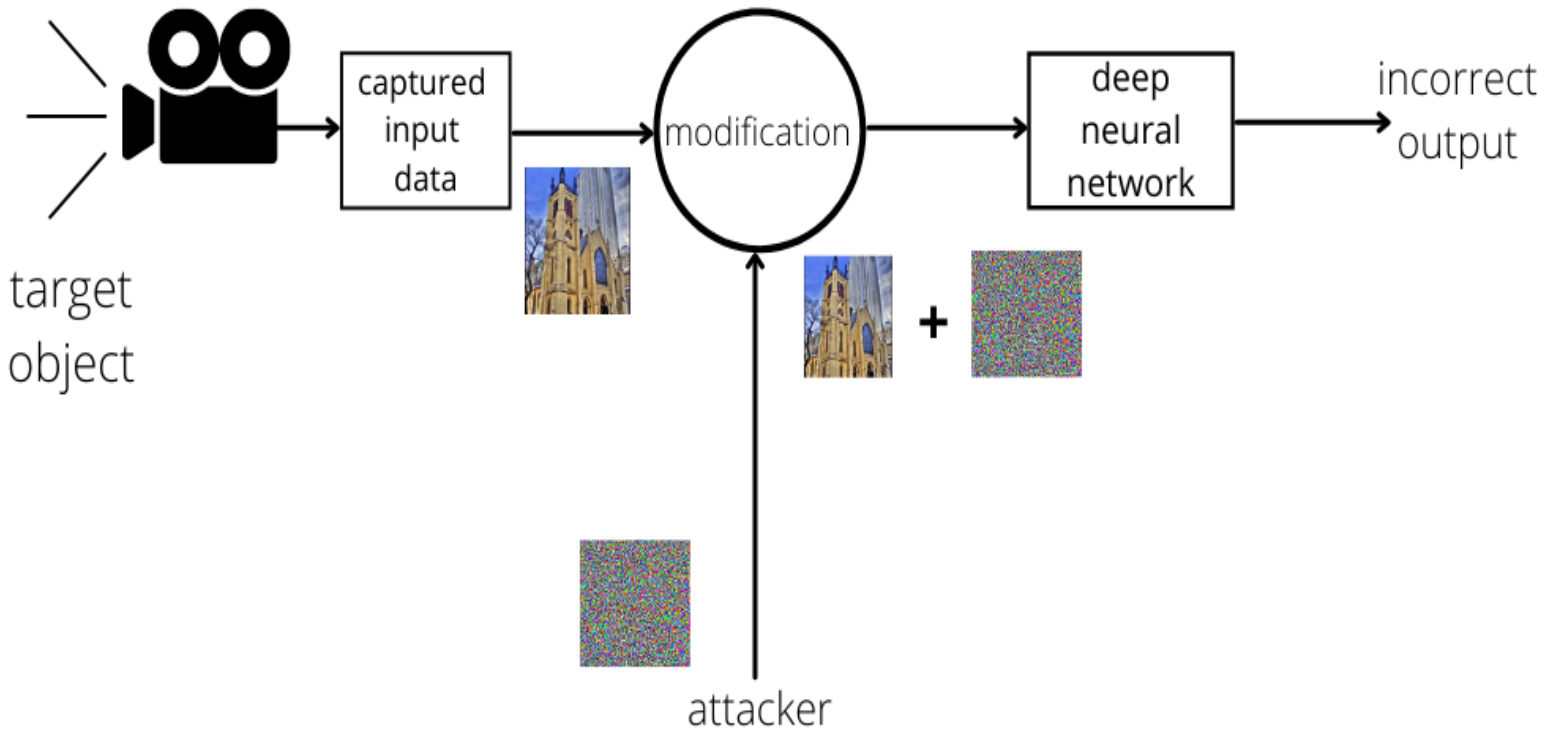



Figure 2: Pipeline of digital-world attack

# Introduction

*Different adversarial attack scenarios*

- **Physical real world attack** - adversarial patch attack [4], dynamic adversarial patch attack [10], shapeshifter attack [7]

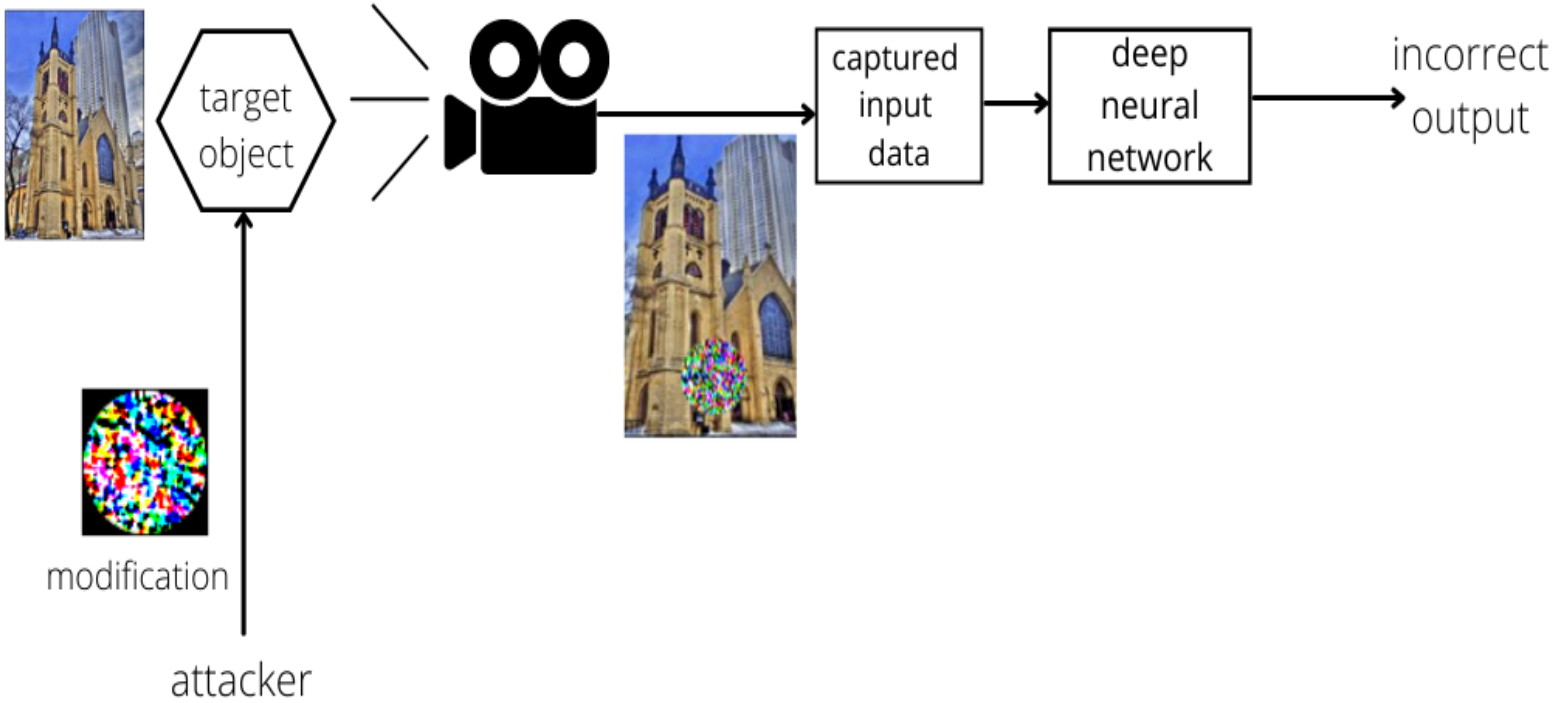



Figure 3: Pipeline of physical real-world attack

# Introduction

*What is the problem?*

- **Adversarial patch attack** - constructing adversarial patch and attaching this patch on the target object to make DNN misclassify it

# Introduction

*What is the problem?*

- **Adversarial patch attack** - constructing adversarial patch and attaching this patch on the target object to make DNN misclassify it


Figure 4: Example of adversarial patch attack. Image source : [4]

# Table of Contents

# Motivation

*Why is it relevant?*

- **Autonomous systems**

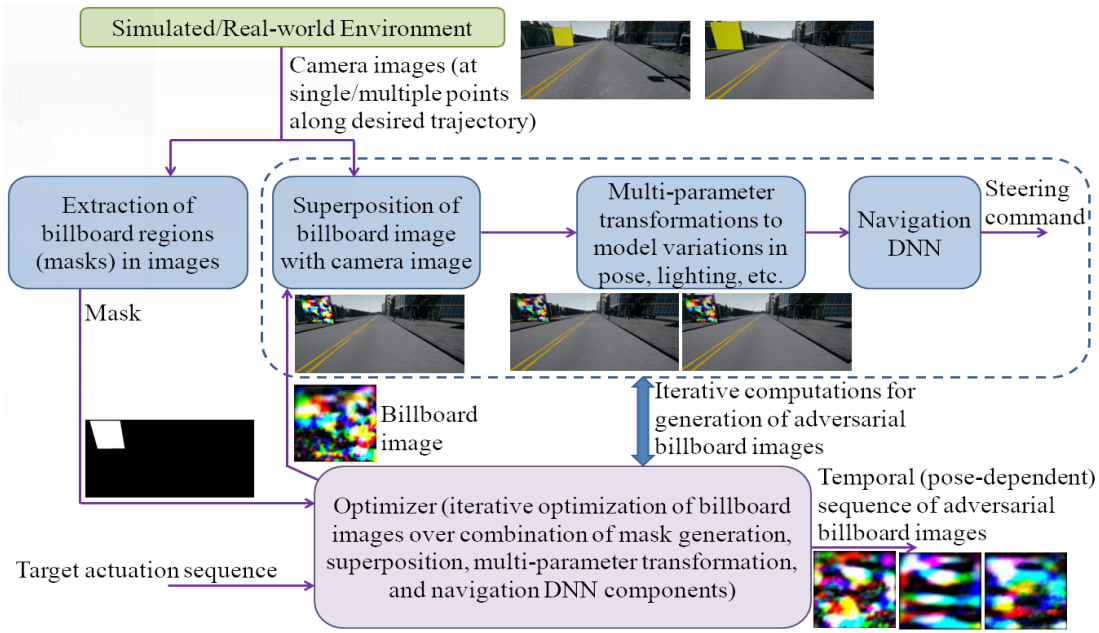



Figure 5: Adversarial attack on autonomous system. Image source: [16]

# Motivation

*Why is it relevant?*

- **Biometrics**

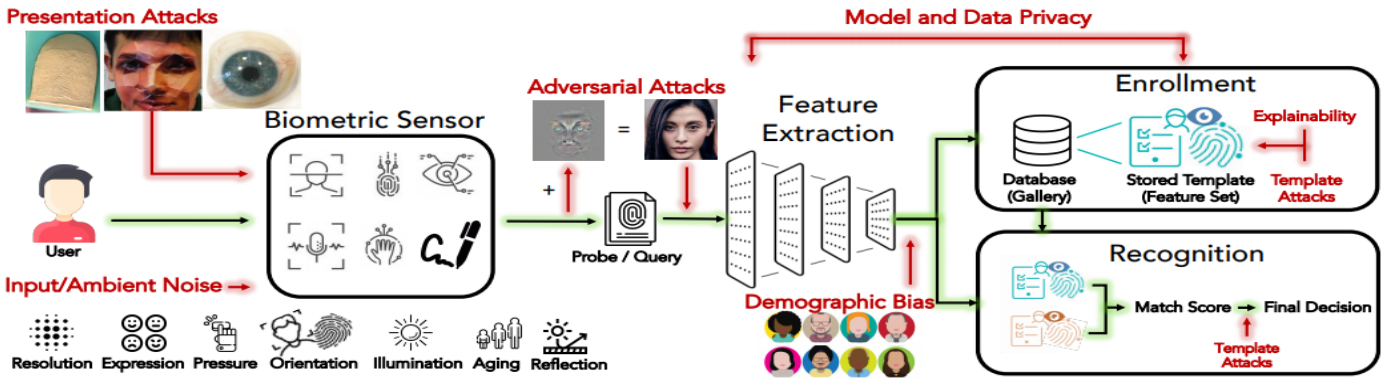



Figure 6: Adversarial attack on biometrics. Image source: [12]

# Motivation

## *Research Scope*

- Intention of helping deep learning research community

# Motivation

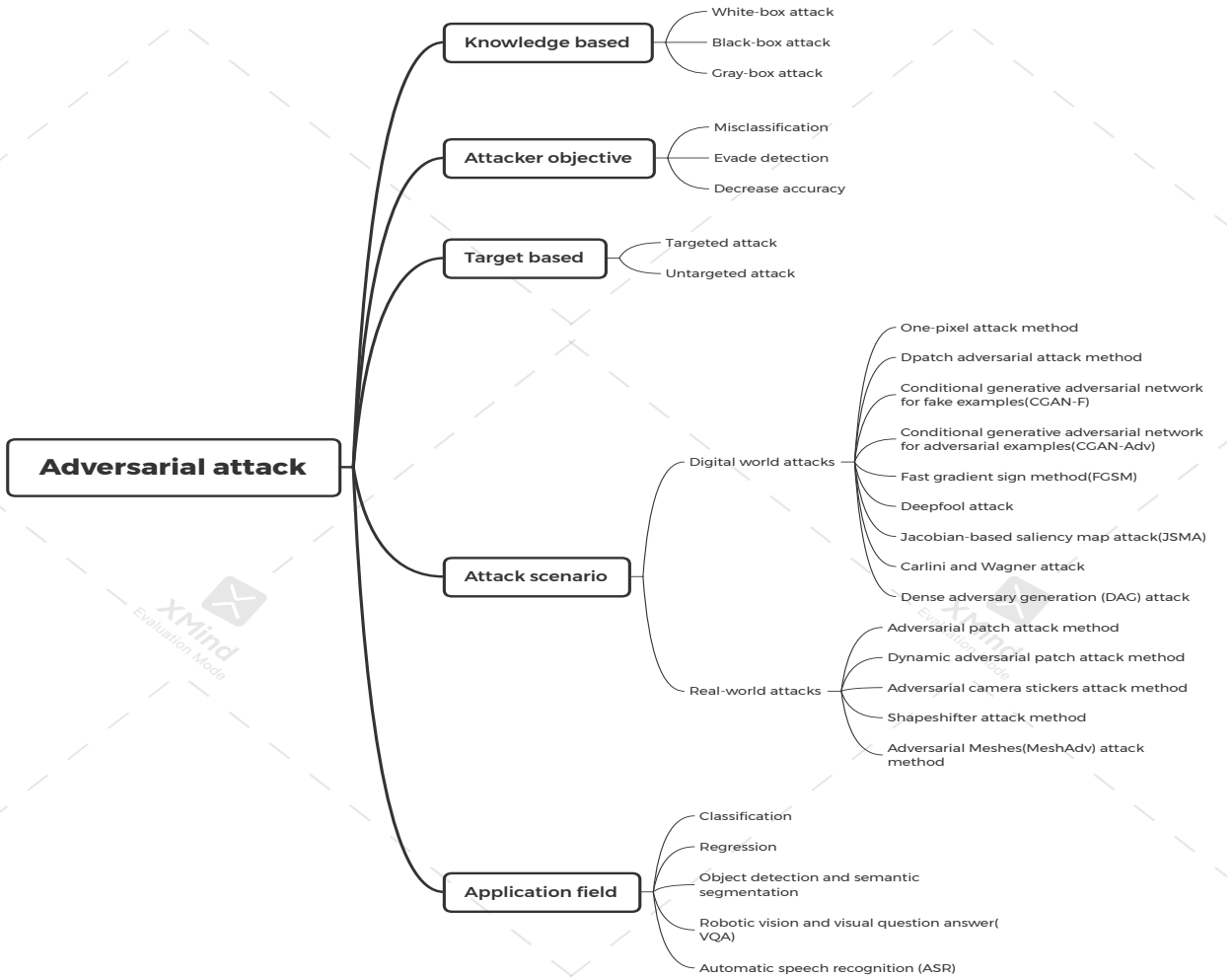
*Research Scope*

- Intention of helping deep learning research community
- Five Research Question (RQ) answered in this project:
  - **RQ1** What are the existing state-of-the-art defense mechanisms against adversarial patch attacks in deep learning?
  - **RQ2** Can an adversarial patch attack always perform the targeted attack?
  - **RQ3** Is adversarial training defense effective against adversarial patch attacks?
  - **RQ4** Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?
  - **RQ5** Can the evidential uncertainty estimation method be used to defend against adversarial patch attacks?

# Table of Contents

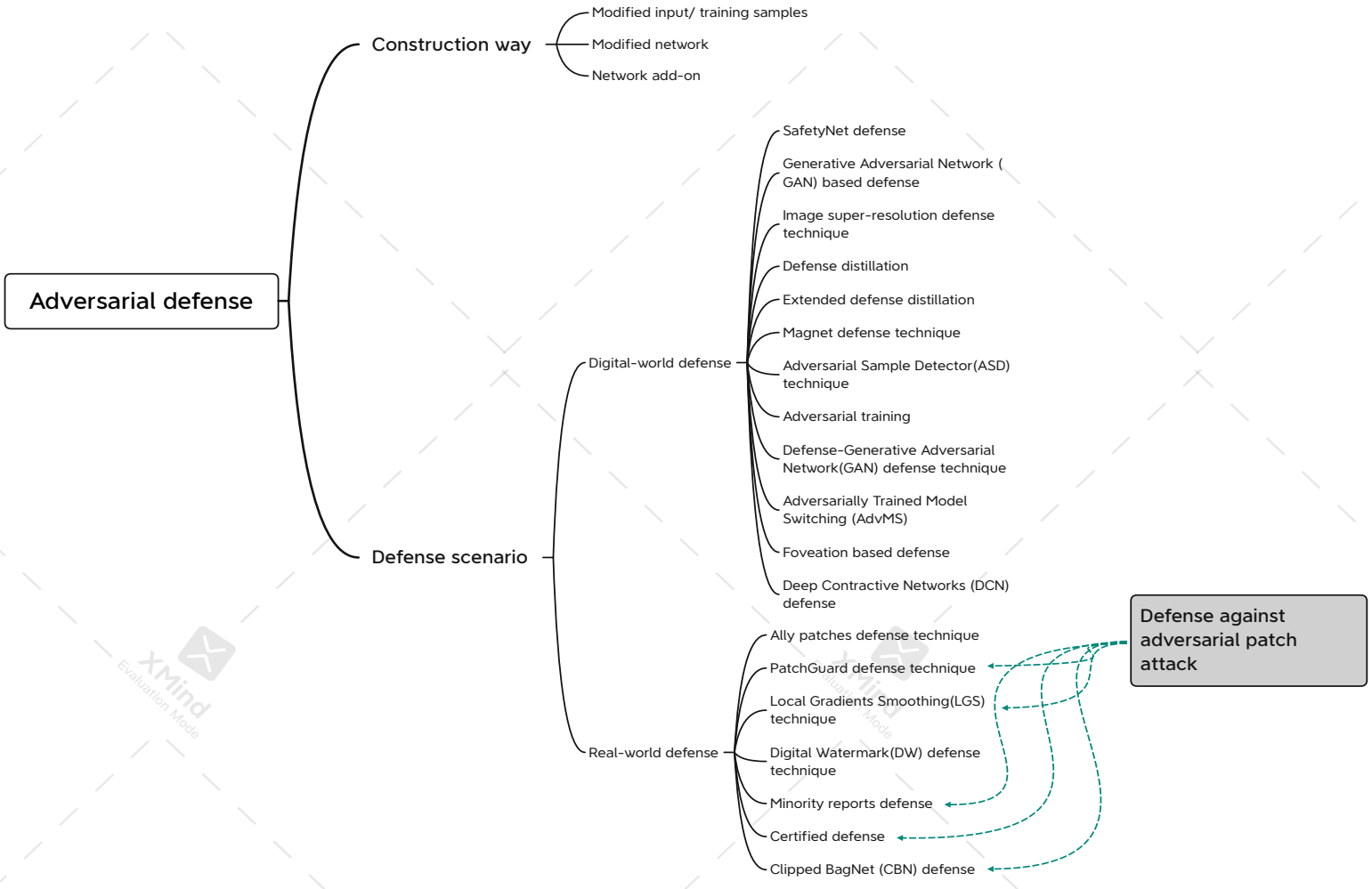
# Related work

*What other people have done?*



# Related work

*RQ1 - What are the existing state-of-the-art defense mechanisms against adversarial patch attacks in deep learning?*



# Table of Contents

# Experimental Setup

## *Datasets*

### **Beans dataset** [14]

Figure 7: Angular Leaf Spot class

Figure 8: Bean Rust class

Figure 9: Healthy class

# Experimental Setup

## *Datasets*

### **Imagenette dataset** [11]

Figure 10: Ball class

Figure 11: Buildings class


Figure 12: Cutting machine class

Figure 13: Dog class

Figure 14: Fish class

# Experimental Setup

*Datasets*

## Imagenette dataset

Figure 15: Musical instrument class

Figure 16: Parachute class


Figure 17: Pump class

Figure 18: Radio class

Figure 19: Truck class

# Experimental Setup

*Datasets*

## RoboCup@Work Dataset

Figure 20: Axis class

Figure 21: Bearing class

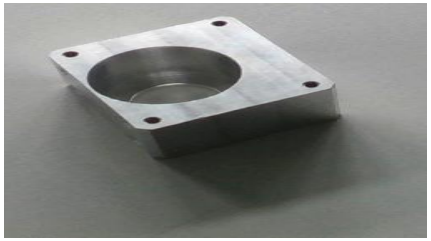


Figure 22: Bearing Box class

Figure 23: Container box blue class

Figure 24: Container box red class

# Experimental Setup

*Datasets*

## RoboCup@Work dataset

Figure 25: Distance Tube class

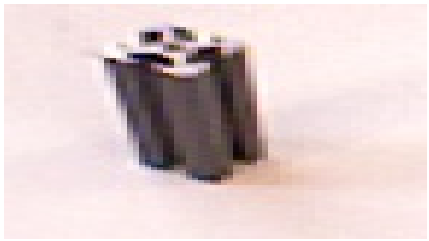


Figure 26: F20_20_B class

Figure 27: F20_20_G class

Figure 28: M20 class

Figure 29: M20_100 class

# Experimental Setup

*Datasets*

## RoboCup@Work dataset

Figure 30: M30 class

Figure 31: Motor class

Figure 32: R20 class

Figure 33: S40_40_B class

Figure 34: S40_40_G class

# Experimental Setup

*Methods*

- MobileNetV2
- ResNet50
- VGG16

# Table of Contents

# Experimental Evaluation

*How adversarial patch attack is performed?*

# Experimental Evaluation

*How adversarial patch attack is performed?*

Adversarial patch generation process

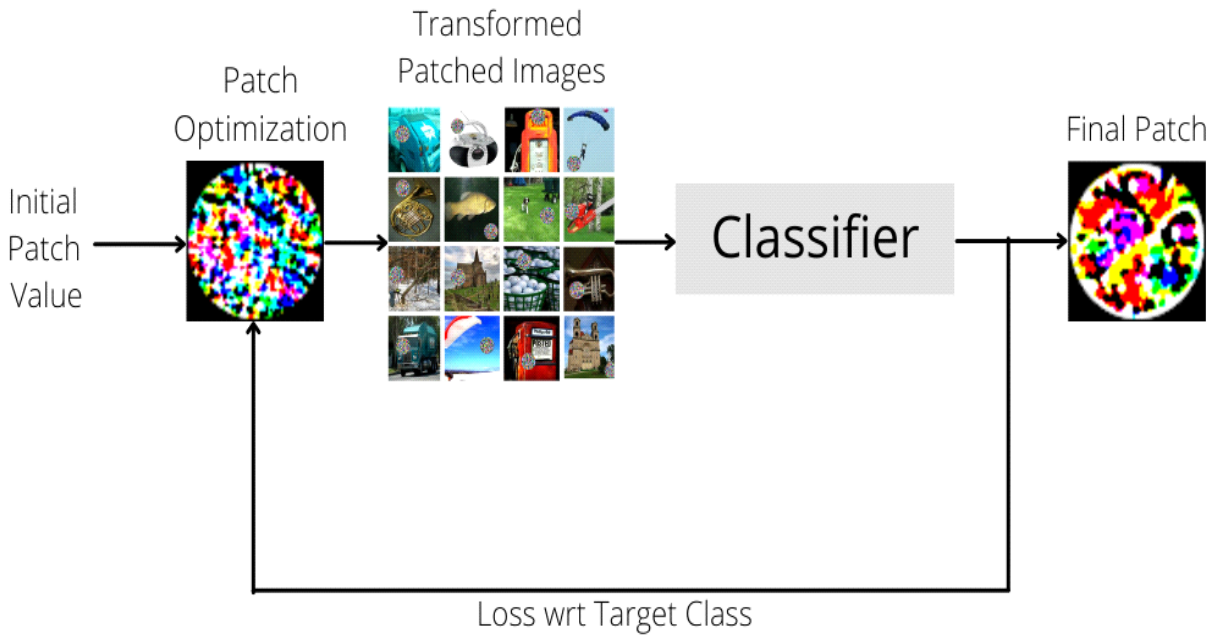



Figure 35: Flow chart of patch generation process

# Experimental Evaluation

*How adversarial patch attack is performed?*

## Adversarial patch attack

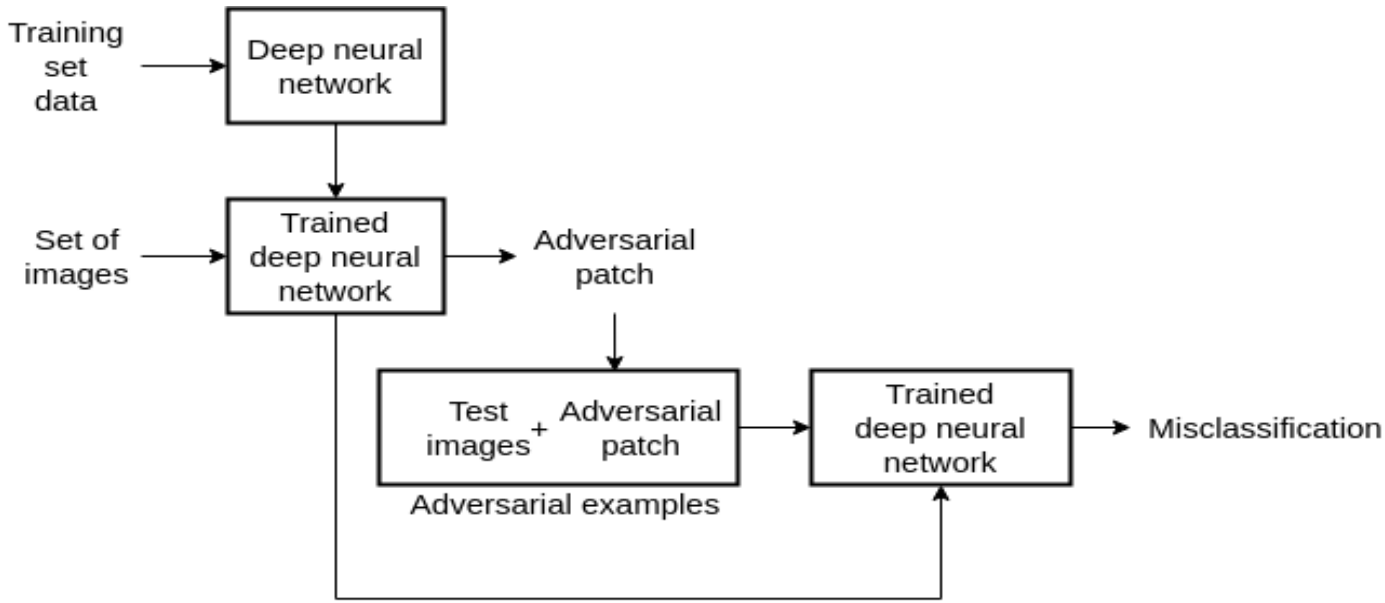



Figure 36: Flow chart of adversarial patch attack

# Experimental Evaluation

*Type of adversarial attack*

- **Non-Targeted Adversarial Attack [9] [17]:** aim is to misclassify the input data as random other label class
- **Targeted Adversarial Attack [13] [5] [22]:** aim is to misclassify the input data as the attacker's desired label class

# Experimental Evaluation

*RQ2 - Can an adversarial patch attack always perform the targeted attack?*

# Experimental Evaluation

*RQ2 - Can an adversarial patch attack always perform the targeted attack?*

**Hypothesis:** an adversarial patch attack can always perform targeted attack

# Experimental Evaluation

*RQ2 - Can an adversarial patch attack always perform the targeted attack?*

**Hypothesis:** an adversarial patch attack can always perform targeted attack

**Observation**: Imagenette dataset

Before targetted adversarial patch attack

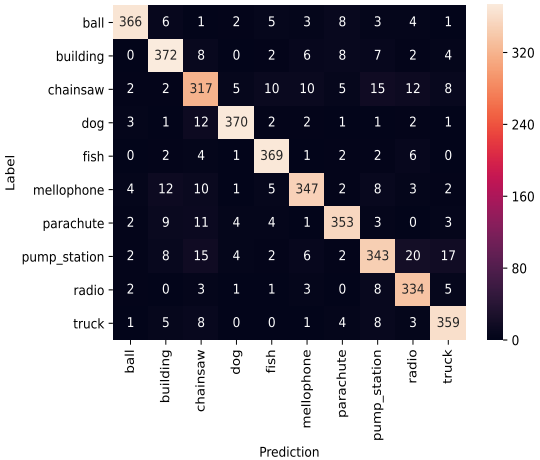



Figure 37: MobileNetV2

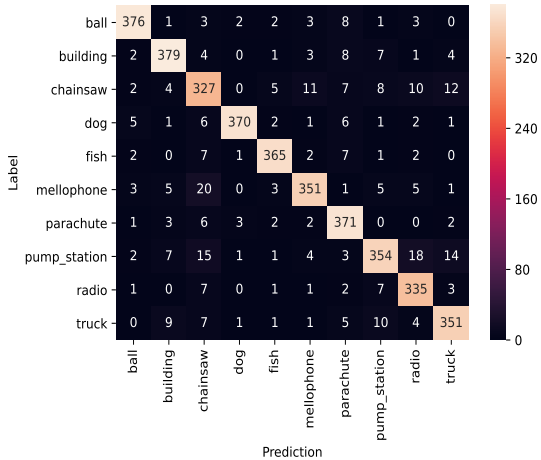



Figure 38: ResNet50

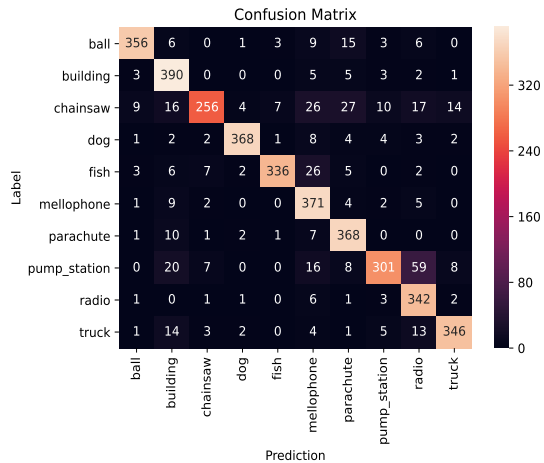
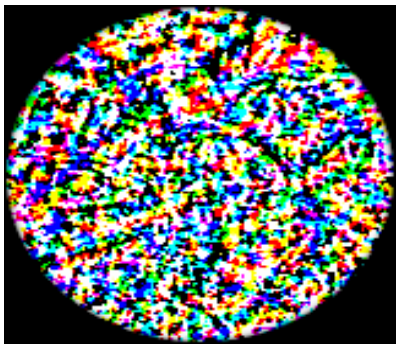


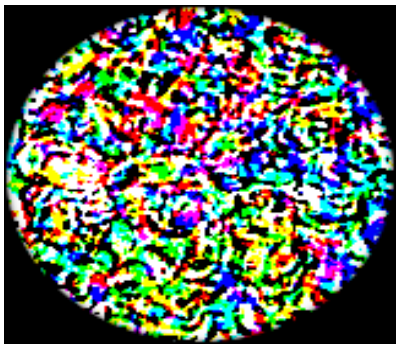

Figure 39: VGG16

# Experiments

*RQ2 - Can an adversarial patch attack always perform the targeted attack?*

Patch generated to perform adversarial patch attack





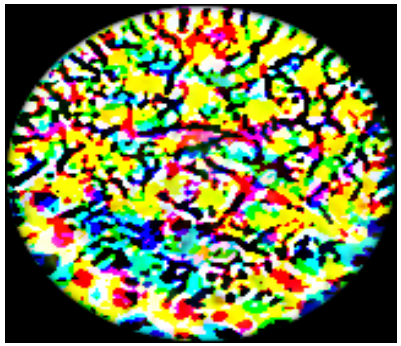


Figure 40: MobileNetV2 Figure 41: ResNet50 Figure 42: VGG16

# Experiments

*RQ2 - Can an adversarial patch attack always perform the targeted attack?*

After targeted adversarial patch attack on ball class

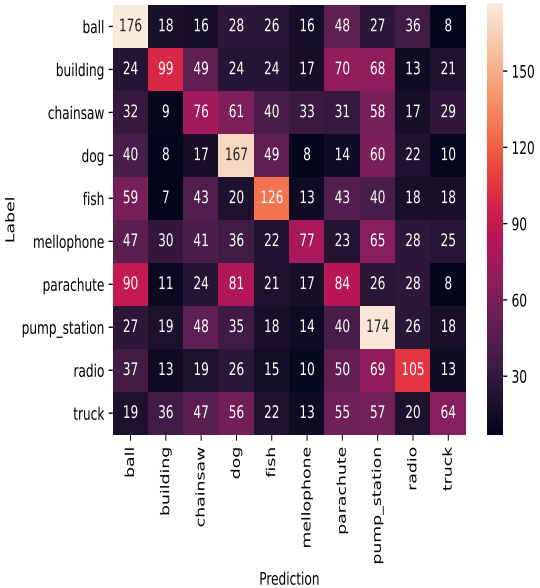



Figure 43: MobileNetV2

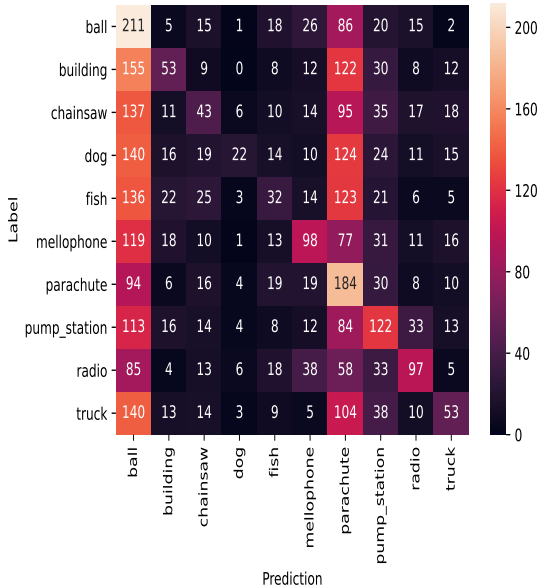



Figure 44: ResNet50

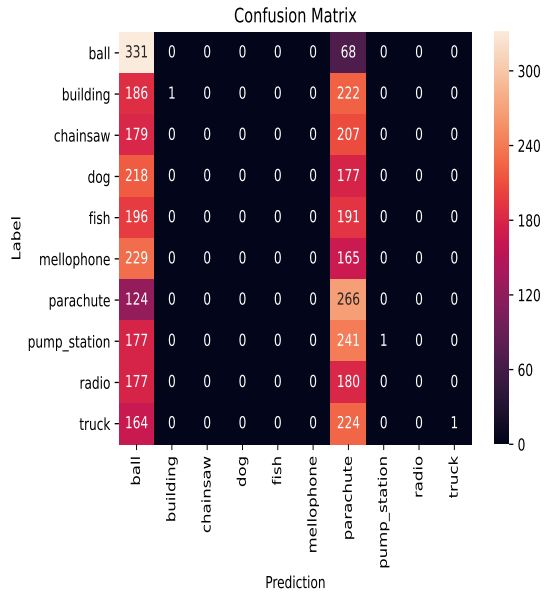



Figure 45: VGG16

# Experiments

*RQ2 - Can an adversarial patch attack always perform the targeted attack?*

After targeted adversarial patch attack on ball class

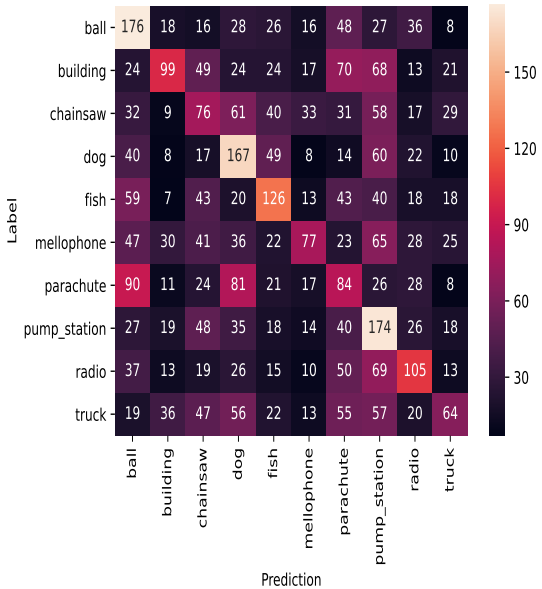



Figure 46: MobileNetV2

# Experiments

*RQ2 - Can an adversarial patch attack always perform the targeted attack?*

After targeted adversarial patch attack on ball class

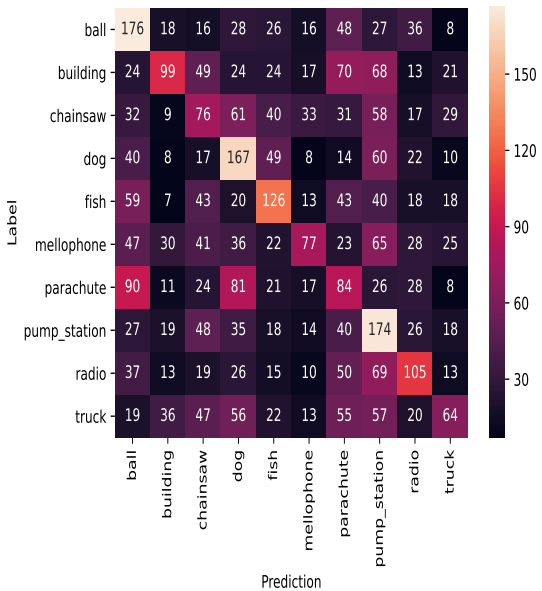



Figure 46: MobileNetV2

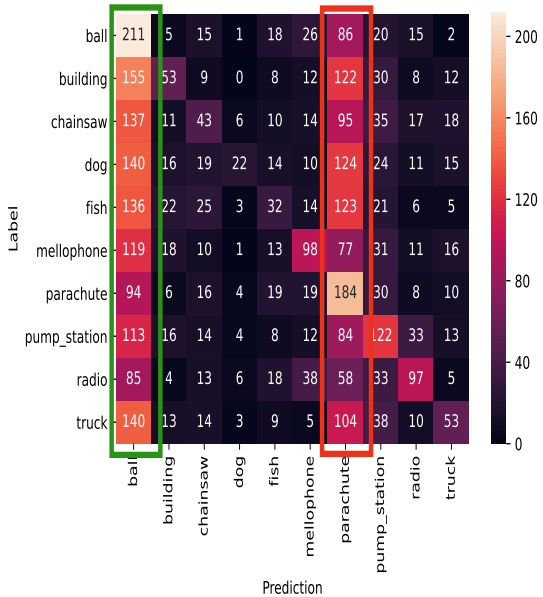



Figure 47: ResNet50

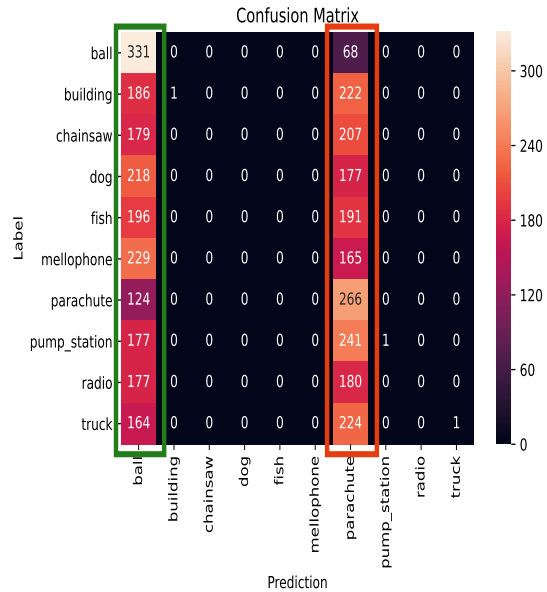



Figure 48: VGG16

# Experiments

## *Defense mechanisms*

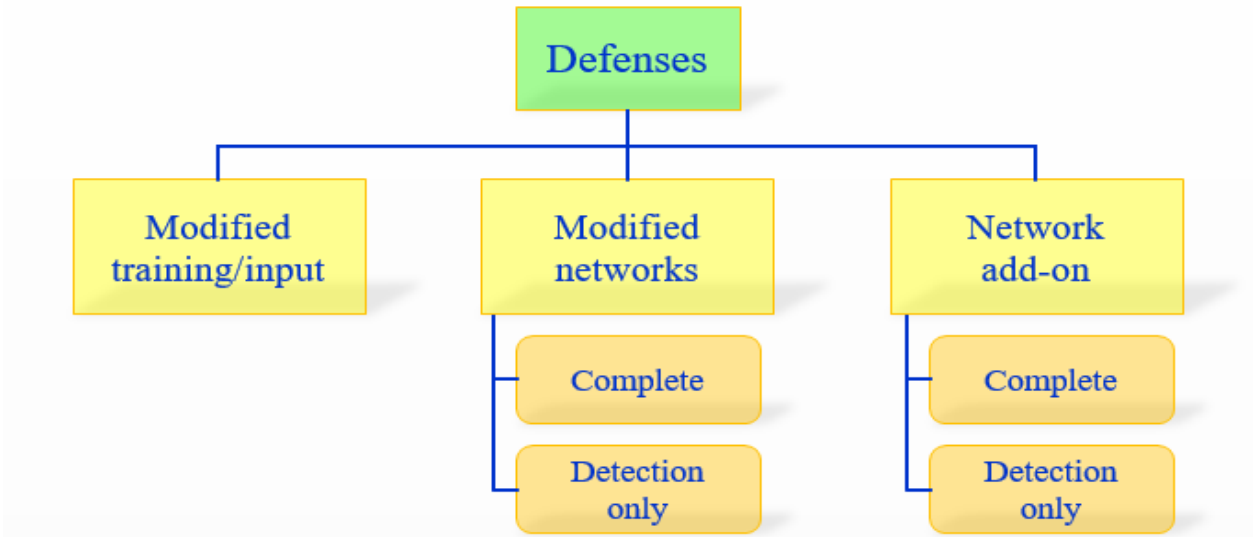



Figure 49: Different ways of categorization adversarial defense mechanisms. Image source: [1]

# Experiments

## *Adversarial training defense*

Adversarial training defense is one of the state-of-the-art defense technique against digital world attacks[3] [19] [15] [2] [24] [23] [20]

# Experiments

## *Adversarial training defense*

Adversarial training defense is one of the state-of-the-art defense technique against digital world attacks[3] [19] [15] [2] [24] [23] [20]
Adversarial training defense process

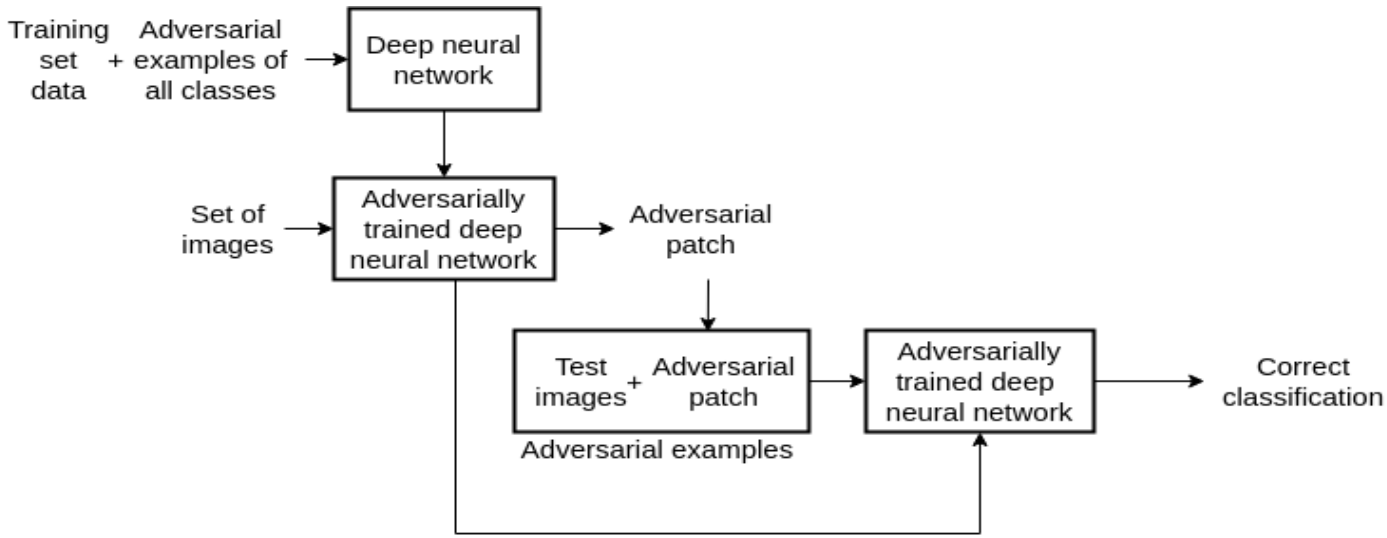



Figure 50: Flowchart of adversarial training defense

# Experiments

*RQ3 - Is adversarial training defense effective against adversarial patch attacks?*

# Experiments

*RQ3 - Is adversarial training defense effective against adversarial patch attacks?*

**Hypothesis:** adversarial training defense is effective against an adversarial patch attack

# Experiments

*RQ3 - Is adversarial training defense effective against adversarial patch attacks?*

**Hypothesis:** adversarial training defense is effective against an adversarial patch attack

Figure 51: Beans dataset

Figure 52: Imagenette dataset

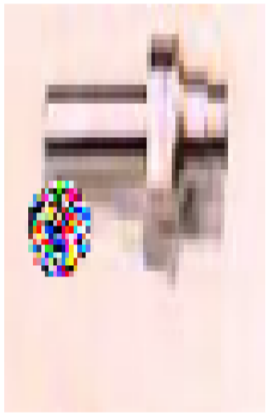


Figure 53: RoboCup@Work dataset

# Experiments

*RQ3 - Is adversarial training defense effective against adversarial patch attacks?*

**Observation:** Imagenette dataset

Before adversarial patch attack

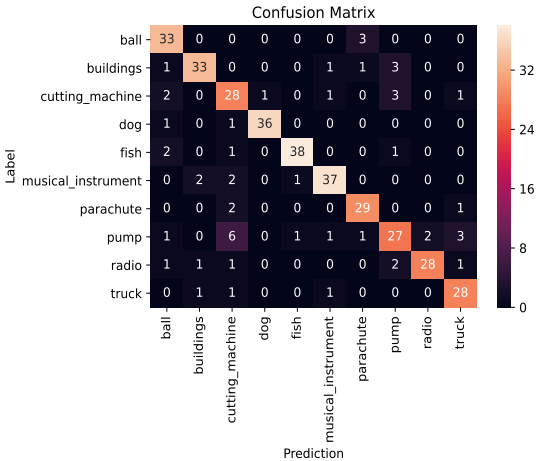



Figure 54: MobileNetV2

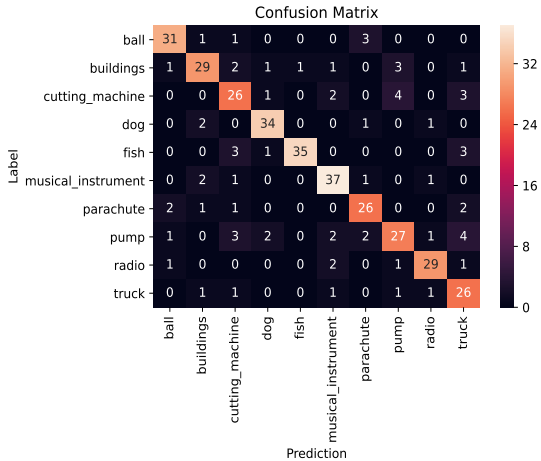



Figure 55: ResNet50

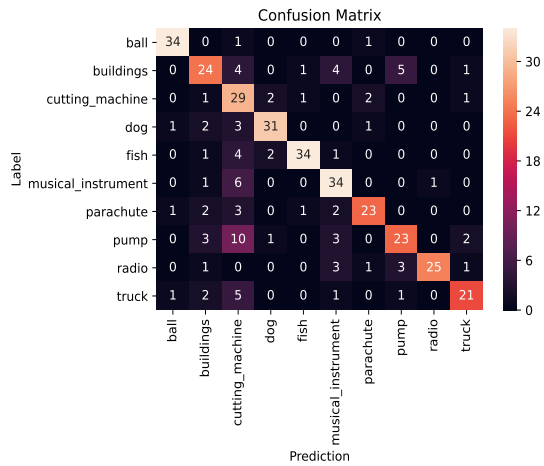



Figure 56: VGG16

# Experiments

*RQ3 - Is adversarial training defense effective against adversarial patch attacks?*

Patch generated to perform adversarial patch attack

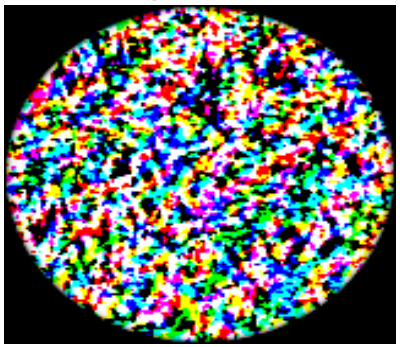


Figure 57: MobileNetV2

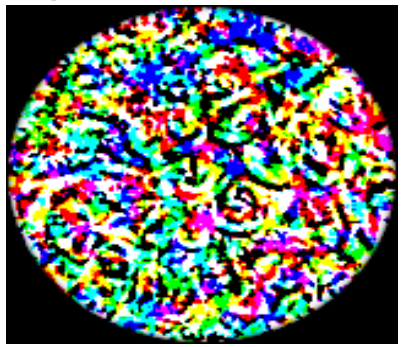


Figure 58: ResNet50

Figure 59: VGG16

# Experiments

*RQ3 - Is adversarial training defense effective against adversarial patch attacks?*

After adversarial patch attack on radio class

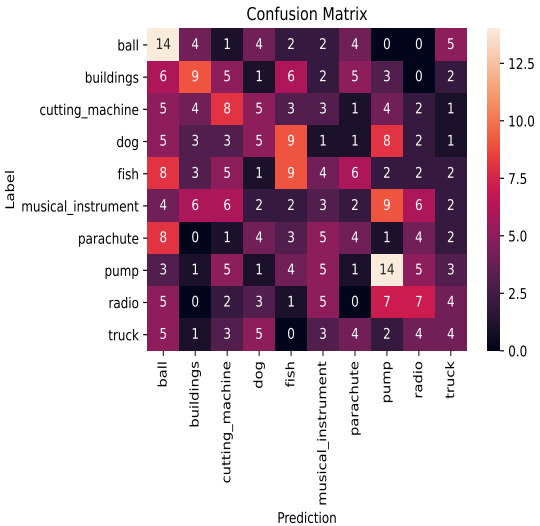



Figure 60: MobileNetV2

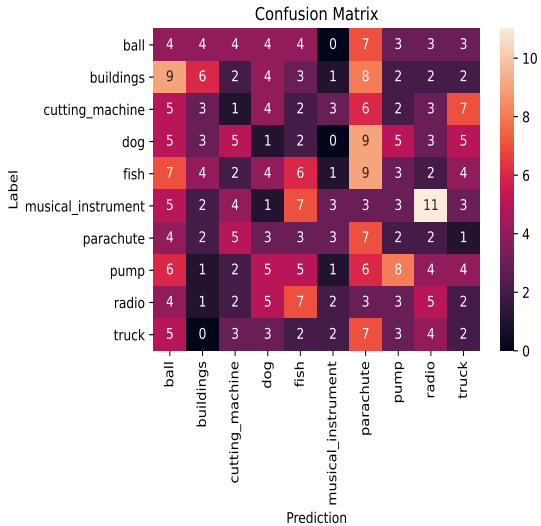



Figure 61: ResNet50

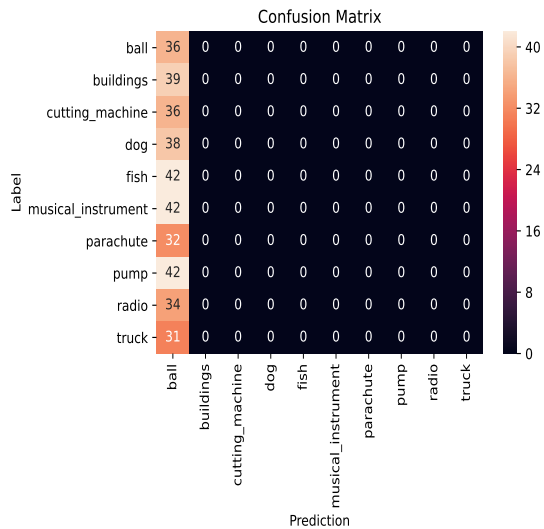



Figure 62: VGG16

# Experiments

*RQ3 - Is adversarial training defense effective against adversarial patch attacks?*

After adversarial patch attack on radio class

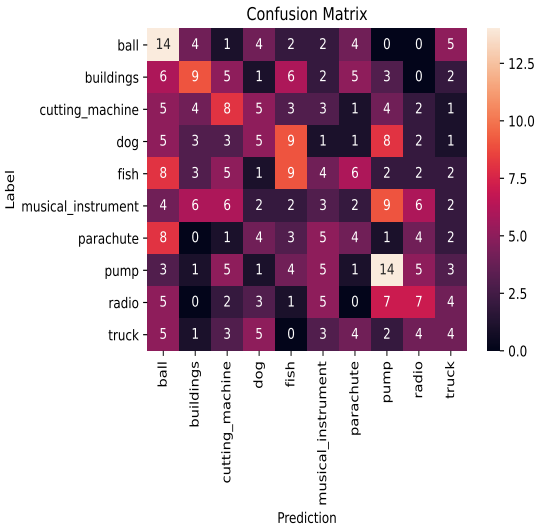



Figure 63: MobileNetV2

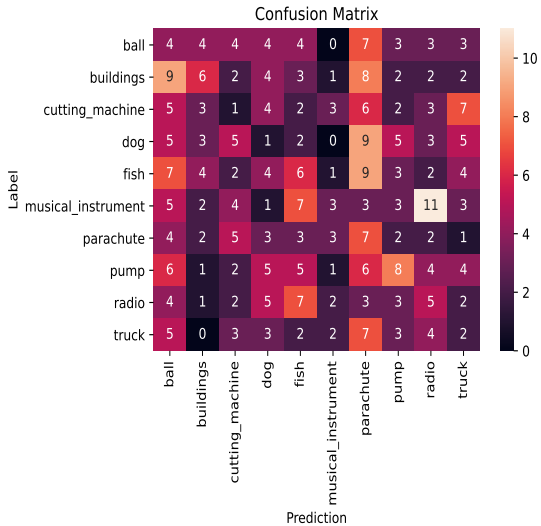



Figure 64: ResNet50

# Experiments

*RQ3 - Is adversarial training defense effective against adversarial patch attacks?*

After adversarial patch attack on radio class

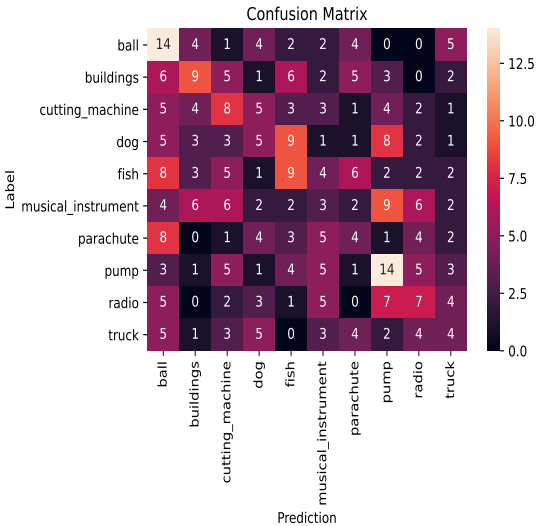



Figure 63: MobileNetV2

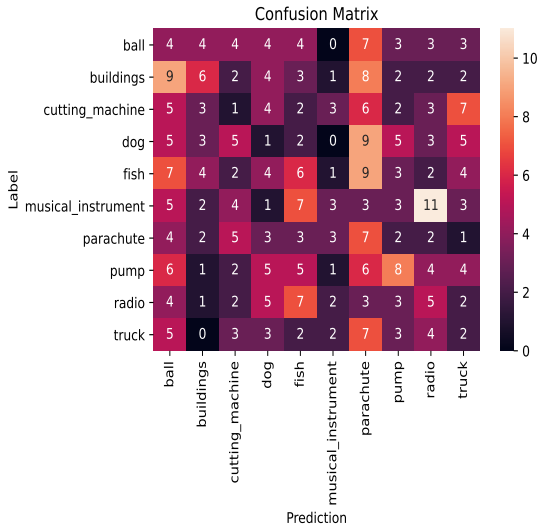



Figure 64: ResNet50

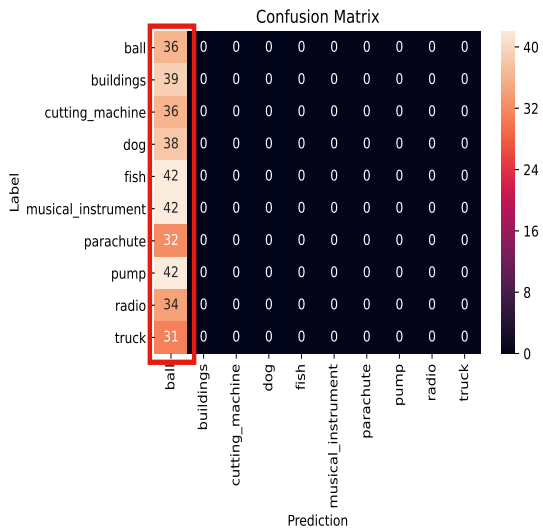



Figure 65: VGG16

# Experiments

*RQ3 - Is adversarial training defense effective against adversarial patch attacks?*

Confidence level of final predictions after adversarial patch attack on radio class

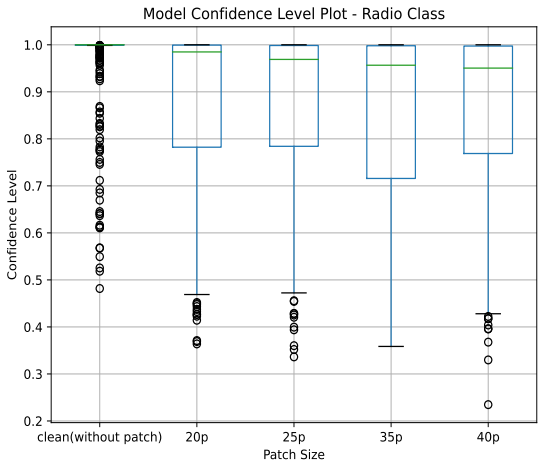



Figure 66: MobileNetV2

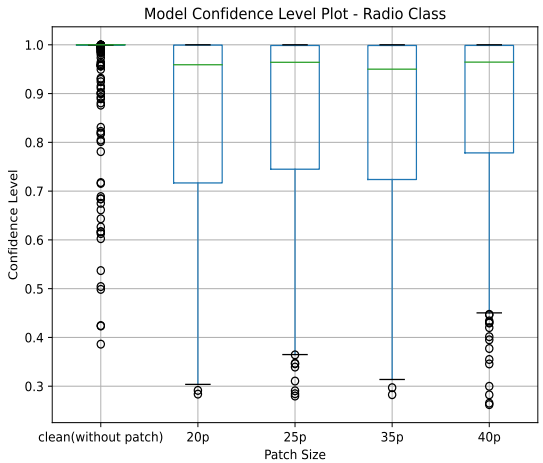



Figure 67: ResNet50

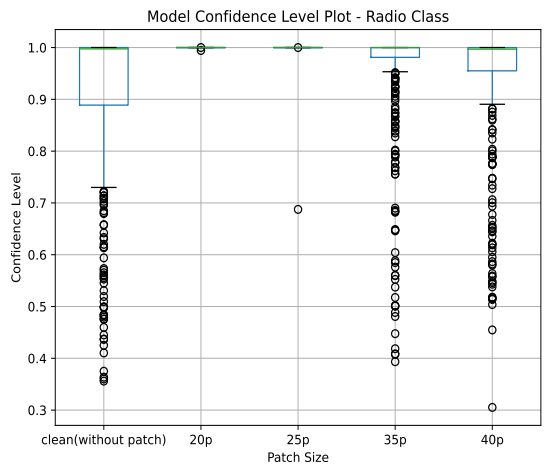



Figure 68: VGG16

# Experiments

*RQ3 - Is adversarial training defense effective against adversarial patch attacks?*

Confidence level of final predictions after adversarial patch attack on radio class

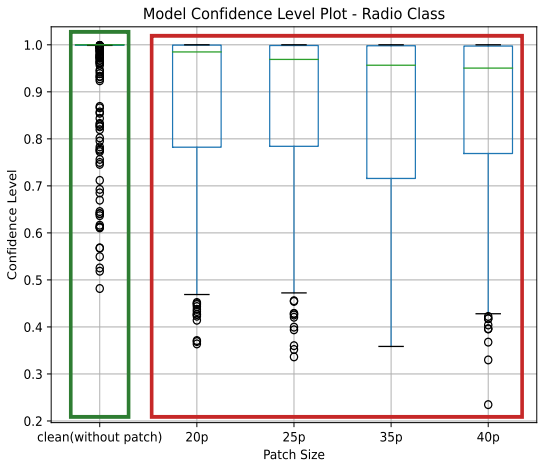



Figure 69: MobileNetV2

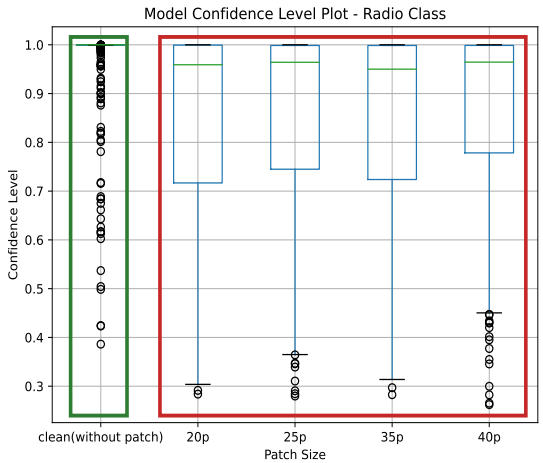



Figure 70: ResNet50

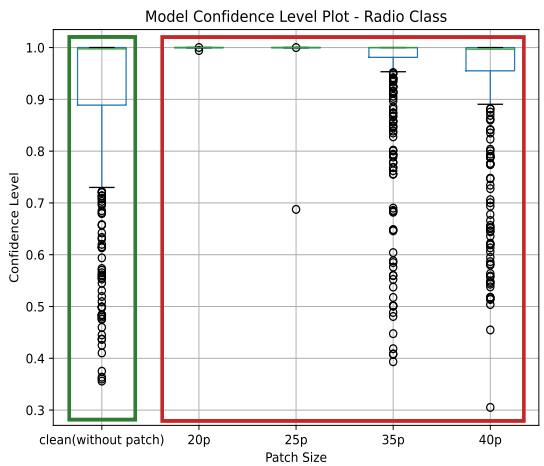



Figure 71: VGG16

# Experiments

## *Abstention class defense*

### Abstention class defense process

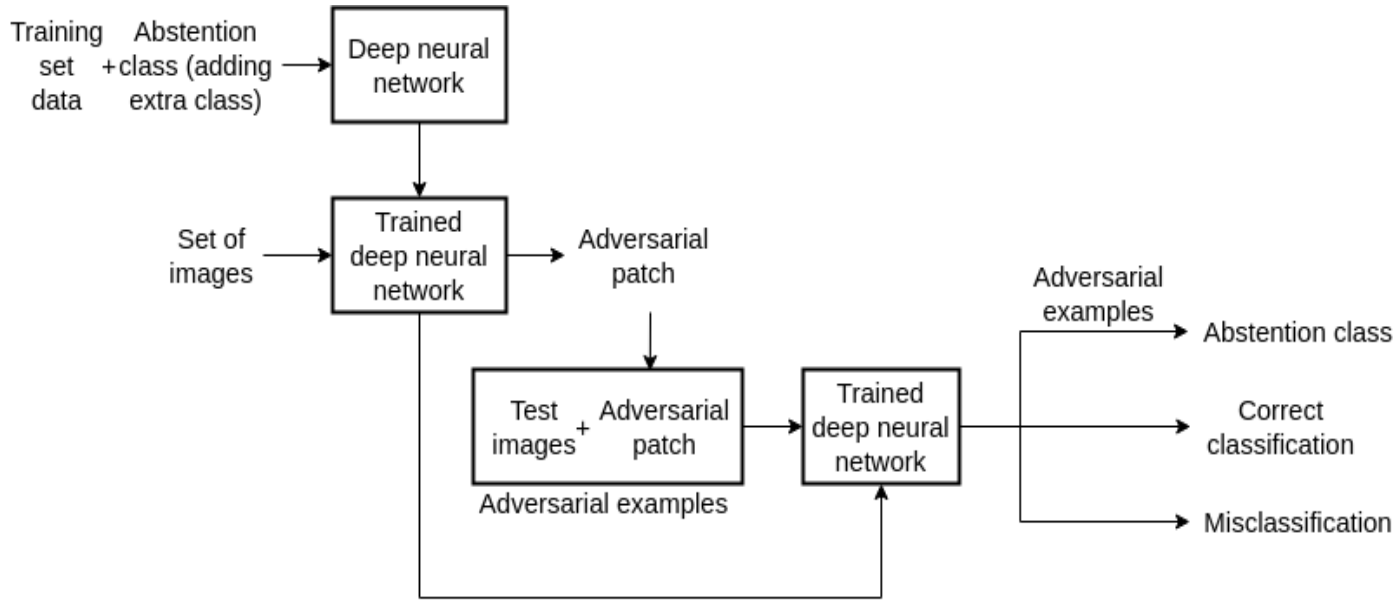



Figure 72: Flowchart of abstention class defense

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

**Hypothesis:** adding abstention class to the training dataset will provide defense against an adversarial patch attack

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

**Hypothesis:** adding abstention class to the training dataset will provide defense against an adversarial patch attack

**Experiment 1**: adding random images to the abstention class

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

**Hypothesis:** adding abstention class to the training dataset will provide defense against an adversarial patch attack
**Experiment 1**: adding random images to the abstention class


Figure 73: Random image 1.
Image source: [25]

Figure 74: Random image 2.
Image source: [27]

Figure 75:
Random image 3.
Image source:
[26]

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

**Experiment 2**: adding adversarial images to the abstention class

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

**Experiment 2**: adding adversarial images to the abstention class

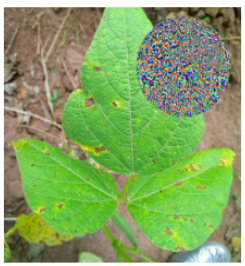


Figure 76: Beans adversarial image

Figure 77: Imagenette adversarial image

Figure 78: RoboCup@Work adversarial image

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

**Experiment 3**: adding subset of all classes present in dataset, to the abstention class

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

**Experiment 3**: adding subset of all classes present in dataset, to the abstention class

**Observation:** Imagenette dataset

Before adversarial patch attack

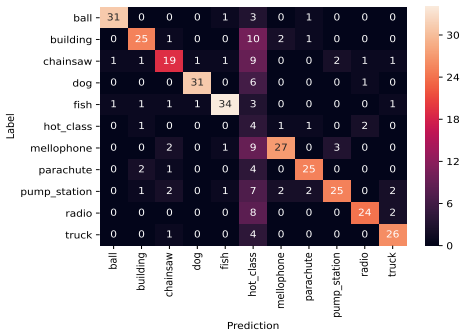



Figure 79: MobileNetV2

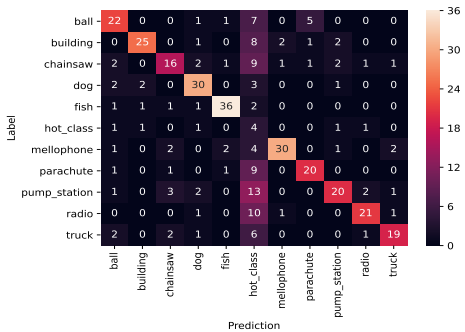



Figure 80: ResNet50

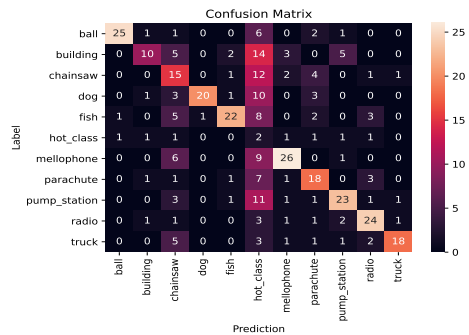



Figure 81: VGG16

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

Patch generated to perform adversarial patch attack

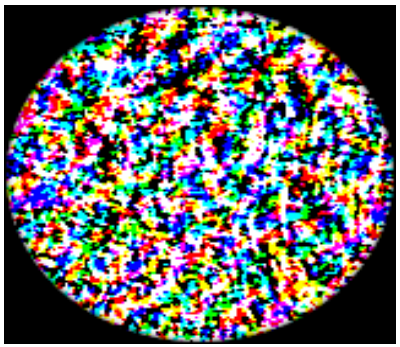


Figure 82: MobileNetV2

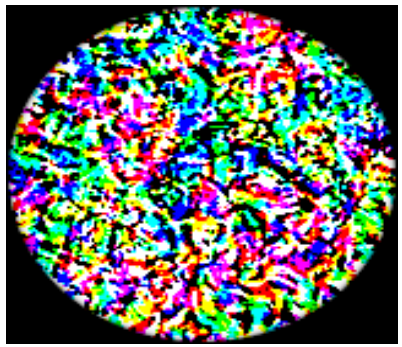


Figure 83: ResNet50

Figure 84: VGG16

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

After adversarial patch attack on mellaphone class

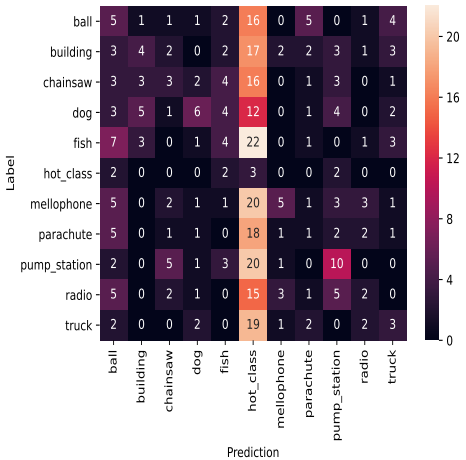



Figure 85: MobileNetV2

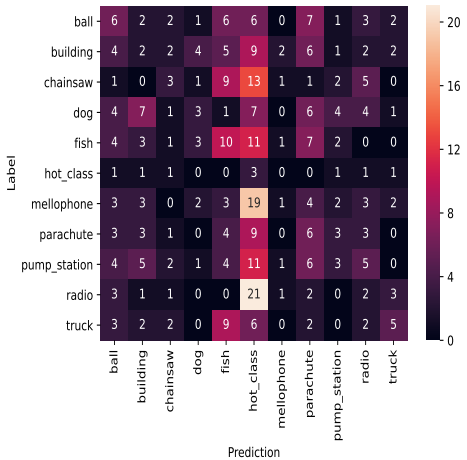



Figure 86: ResNet50

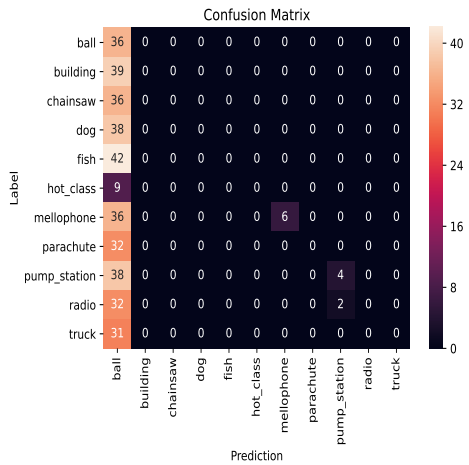



Figure 87: VGG16

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

After adversarial patch attack on mellaphone class

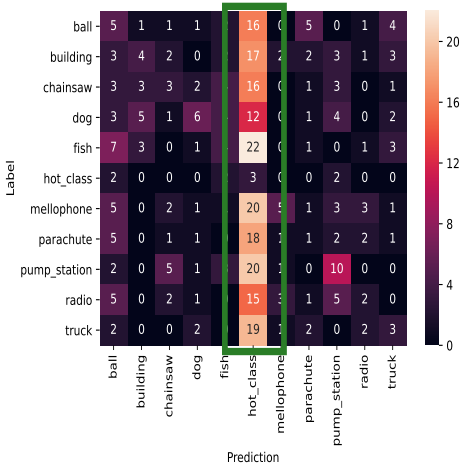



Figure 88: MobileNetV2

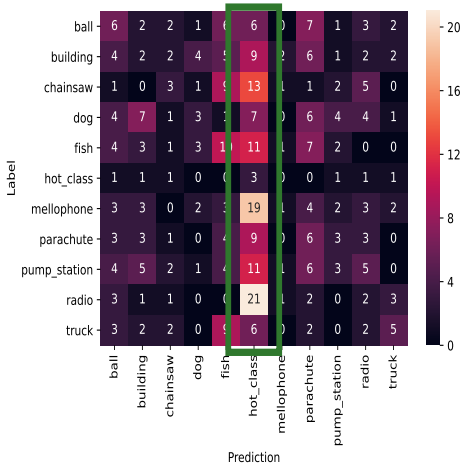



Figure 89: ResNet50

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

After adversarial patch attack on mellaphone class

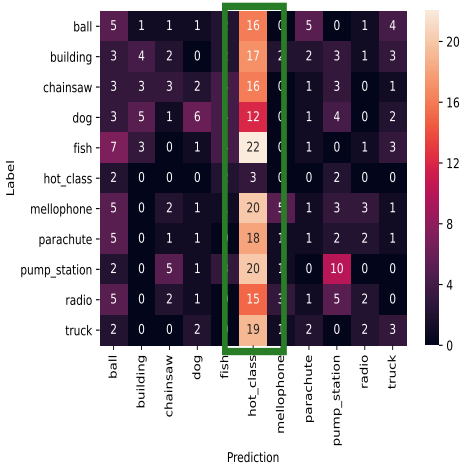



Figure 88: MobileNetV2

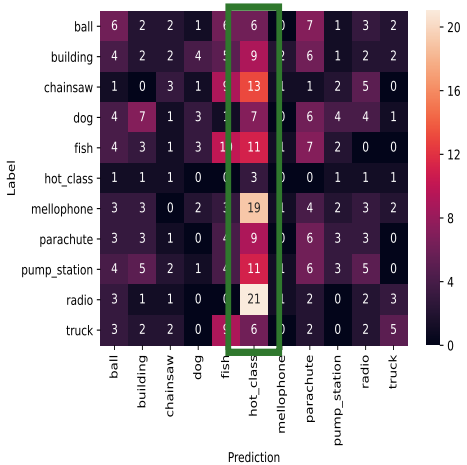



Figure 89: ResNet50

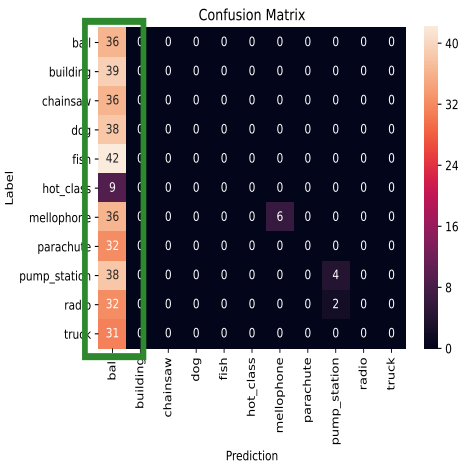



Figure 90: VGG16

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

Confidence level of final predictions after adversarial patch attack on mellaphone class

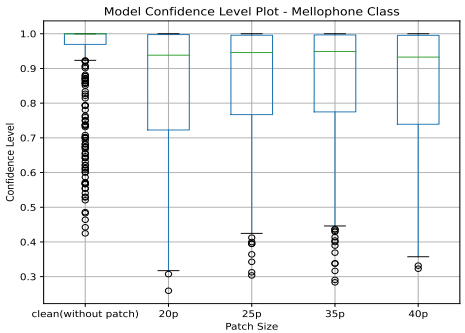



Figure 91: MobileNetV2

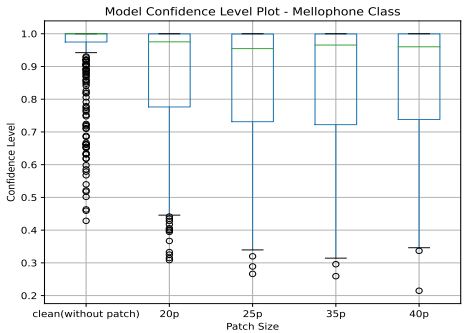



Figure 92: ResNet50

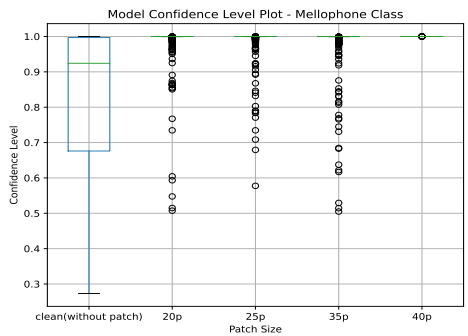



Figure 93: VGG16

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

Confidence level of final predictions after adversarial patch attack on mellaphone class

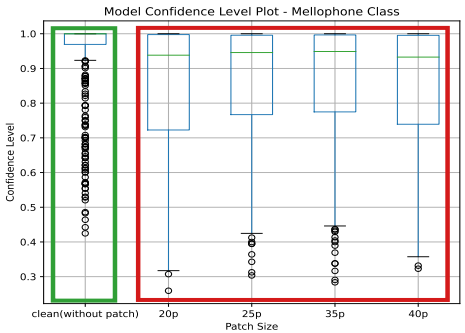



Figure 94: MobileNetV2

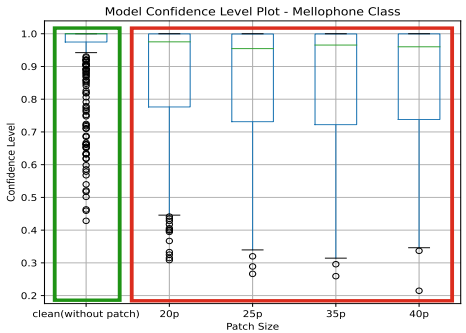



Figure 95: ResNet50

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

Confidence level of final predictions after adversarial patch attack on mellaphone class

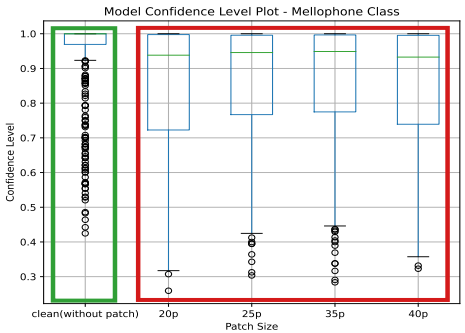



Figure 94: MobileNetV2

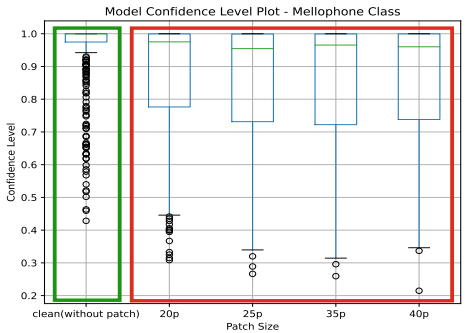



Figure 95: ResNet50

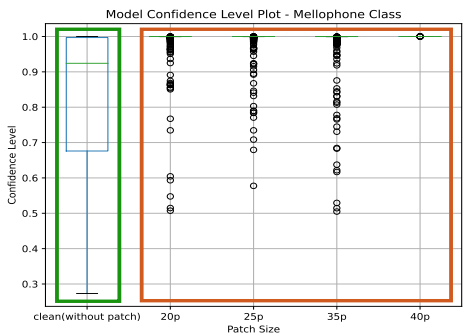



Figure 96: VGG16

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

**Observation:** RoboCup@Work dataset

Before adversarial patch attack

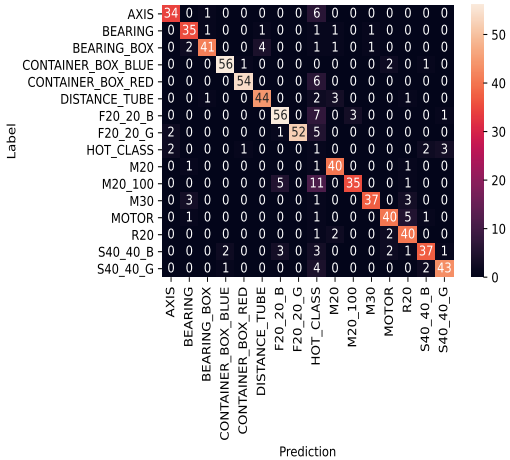



Figure 97: MobileNetV2

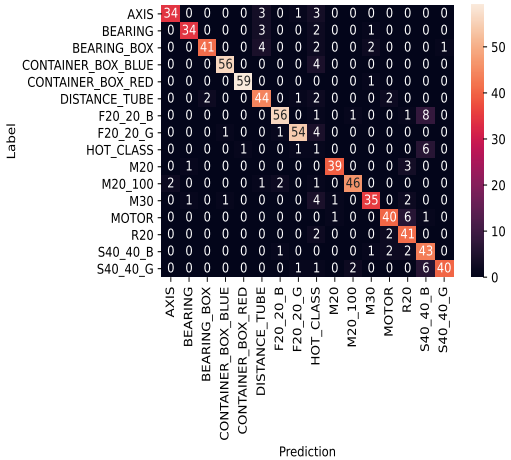



Figure 98: ResNet50

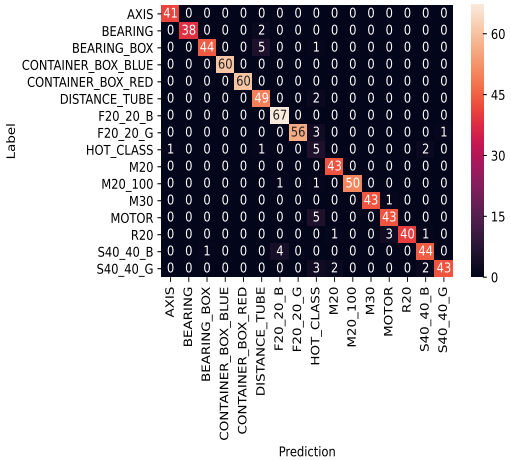



Figure 99: VGG16

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

Patch generated to perform adversarial patch attack

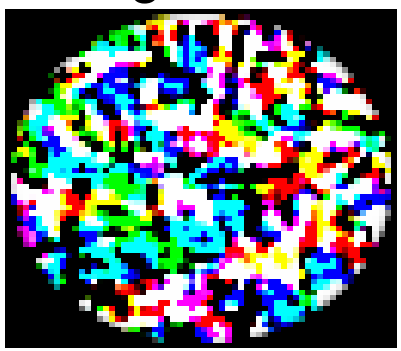


Figure 100: MobileNetV2

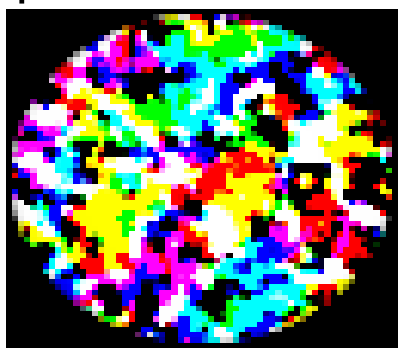


Figure 101: ResNet50

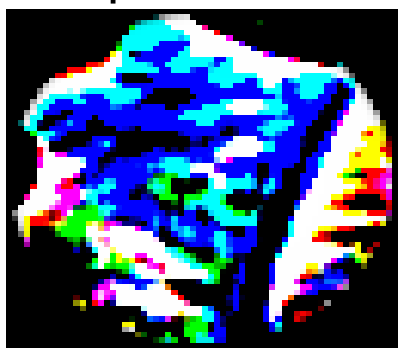


Figure 102: VGG16

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

After adversarial patch attack on M20_100 class

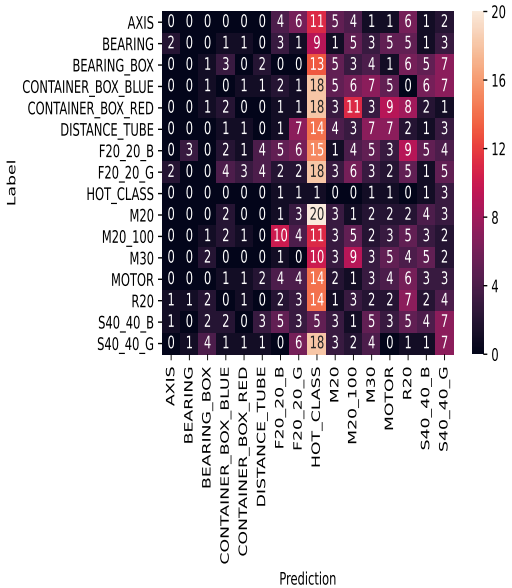



Figure 103: MobileNetV2

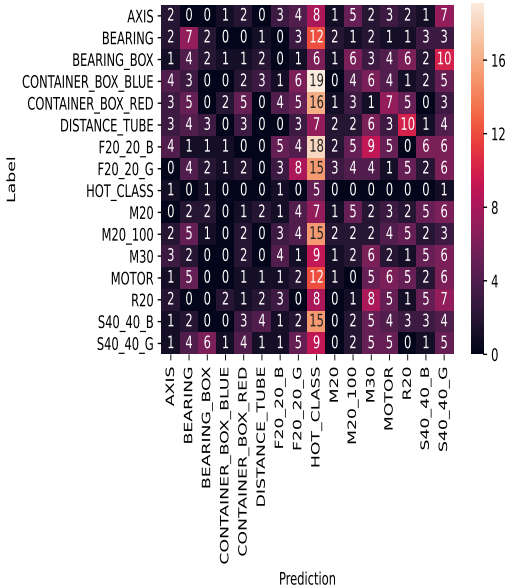



Figure 104: ResNet50

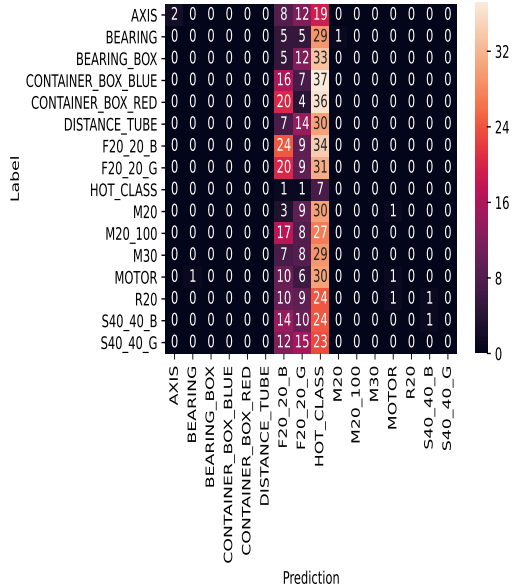



Figure 105: VGG16

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

After adversarial patch attack on M20_100 class

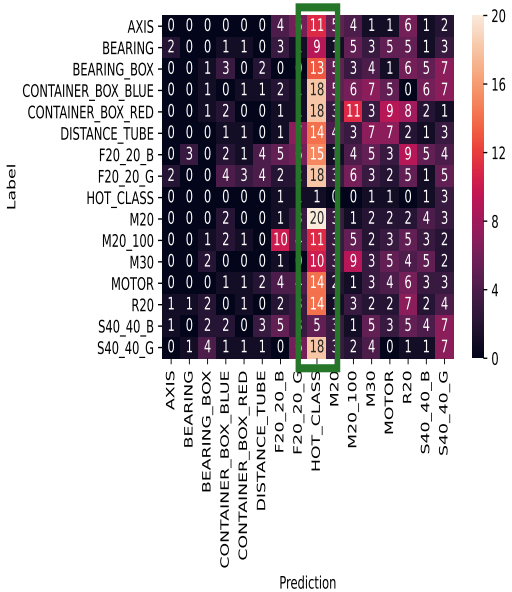



Figure 106: MobileNetV2

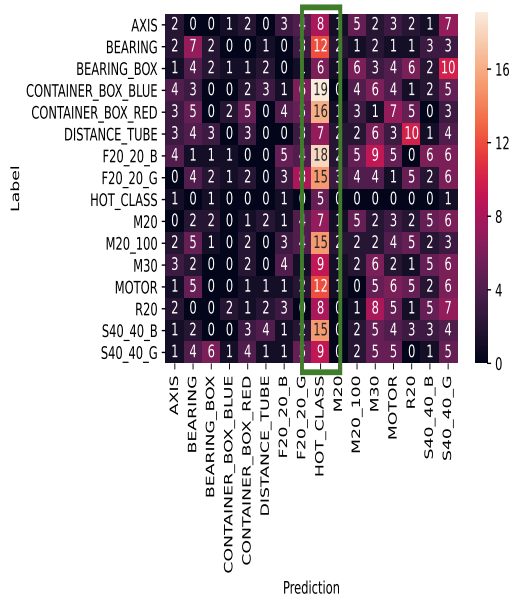



Figure 107: ResNet50

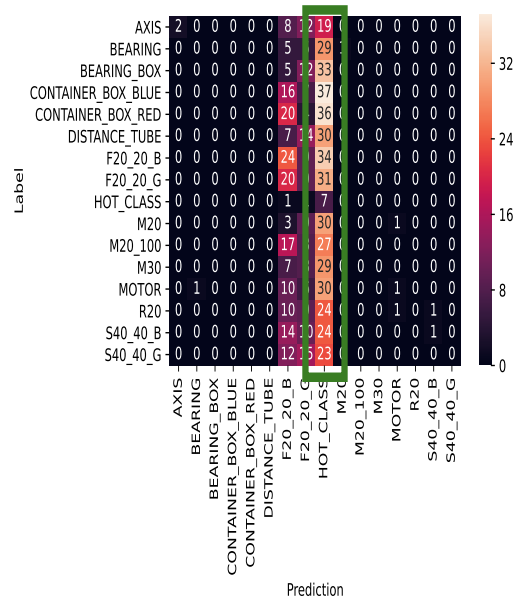



Figure 108: VGG16

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

Confidence level of final predictions after adversarial patch attack on M20_100 class

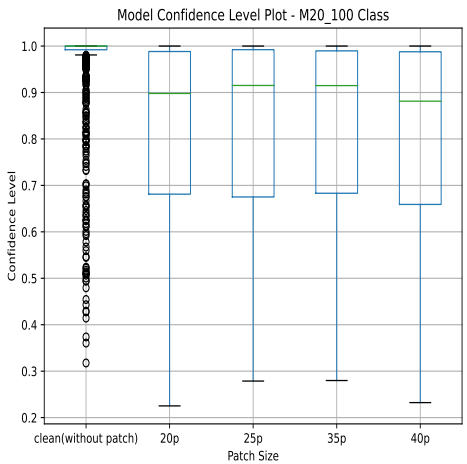



Figure 109: MobileNetV2

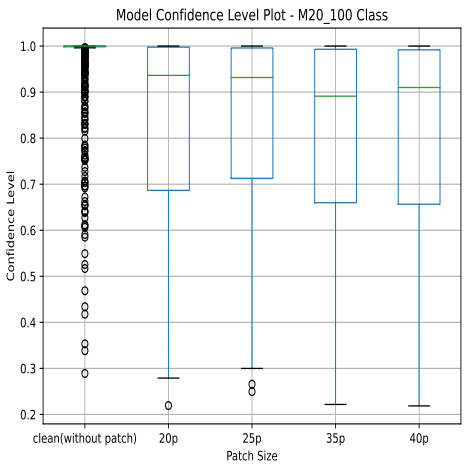



Figure 110: ResNet50

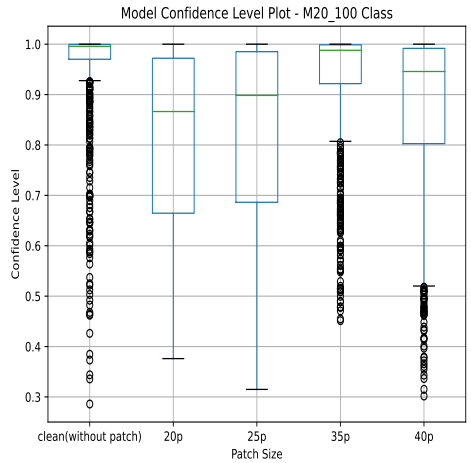



Figure 111: VGG16

# Experiments

*RQ4 - Could the defense against adversarial patch attacks be improved by adding an abstention class to a dataset?*

Confidence level of final predictions after adversarial patch attack on M20_100 class

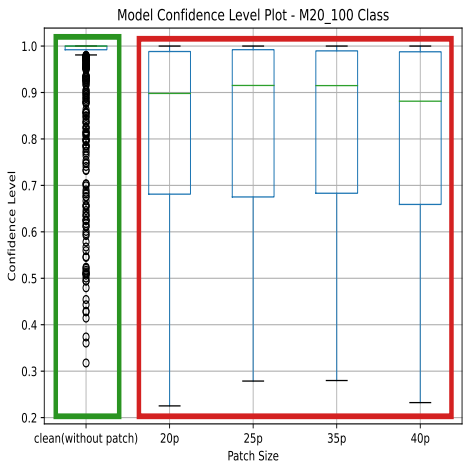



Figure 112: MobileNetV2

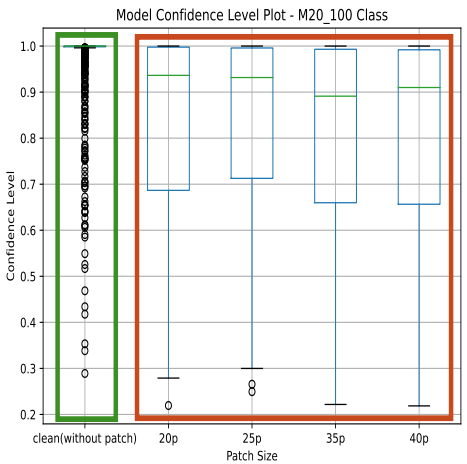



Figure 113: ResNet50

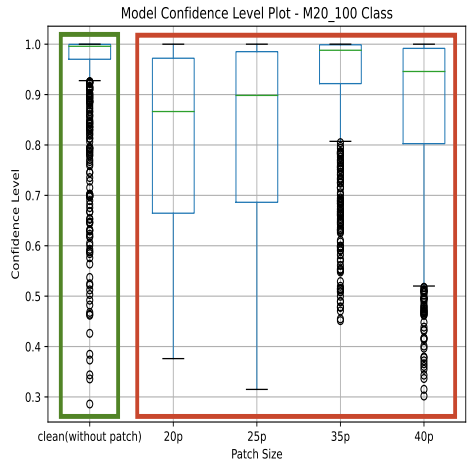



Figure 114: VGG16

# Experiments

## *Evidential uncertainty estimation defense*

Evidential uncertainty estimation defense process

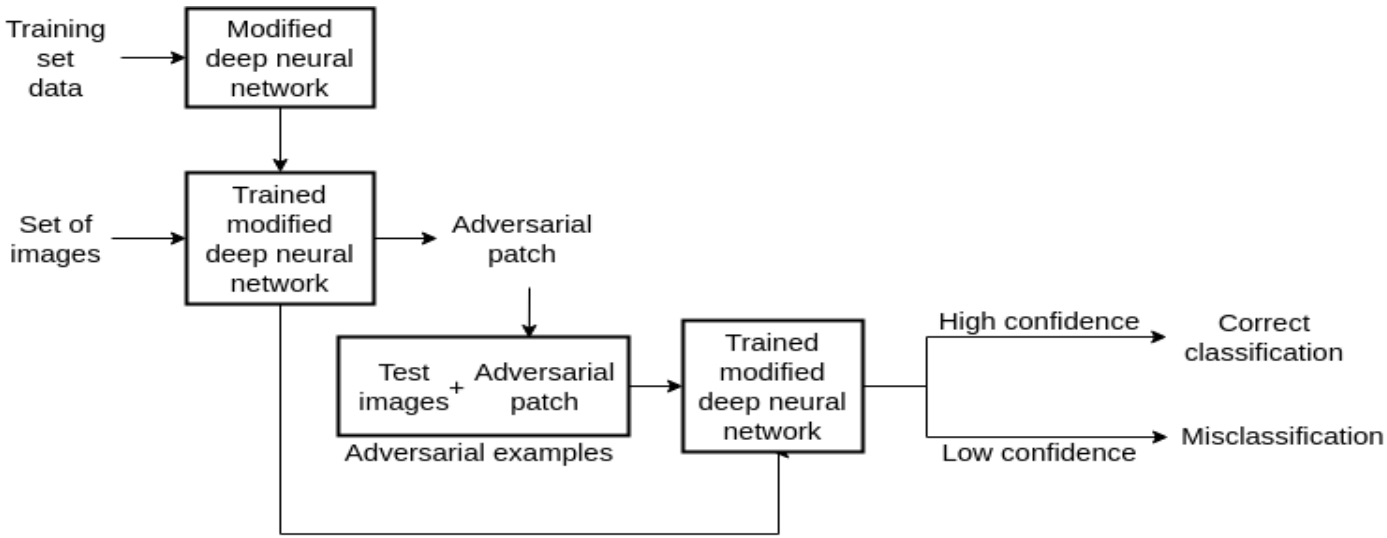



Figure 115: Flowchart of evidential uncertainty estimation defense

# Experiments

*RQ5 - Can the evidential uncertainty estimation method be used to defend against adversarial patch attacks?*

# Experiments

*RQ5 - Can the evidential uncertainty estimation method be used to defend against adversarial patch attacks?*

**Hypothesis:** evidential uncertainty estimation can be used to provide defense against adversarial patch attack

# Experiments

*RQ5 - Can the evidential uncertainty estimation method be used to defend against adversarial patch attacks?*

**Hypothesis:** evidential uncertainty estimation can be used to provide defense against adversarial patch attack

**Observation:** Imagenette dataset

Confusion matrix before adversarial patch attack

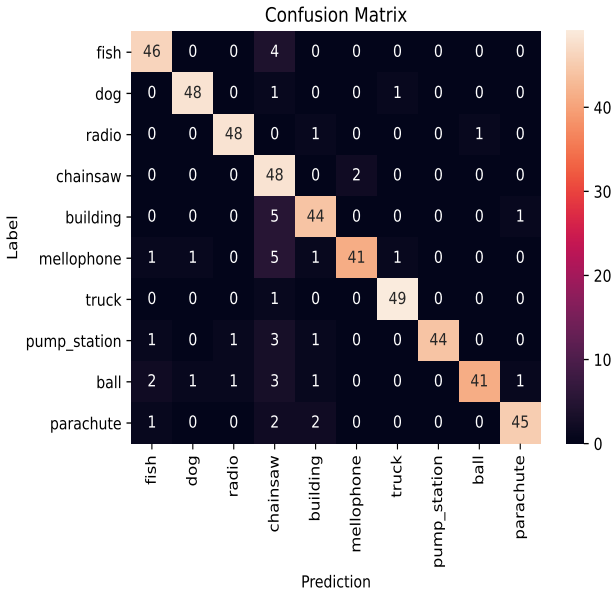



Figure 116: MobileNetV2

# Experiments

*RQ5 - Can the evidential uncertainty estimation method be used to defend against adversarial patch attacks?*

Patch generated to perform adversarial patch attack

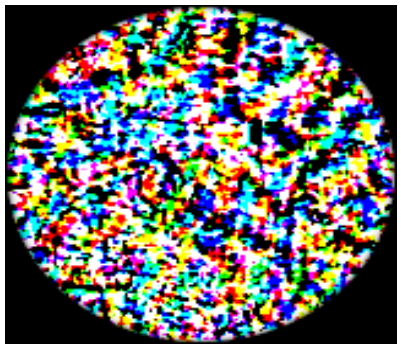


Figure 117: radio class

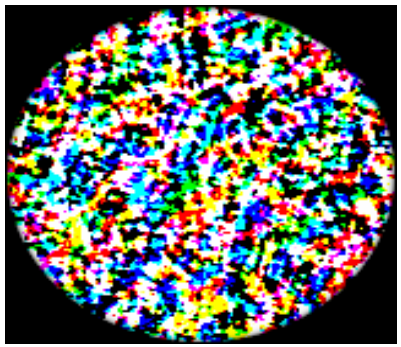


Figure 118: mellaphone class

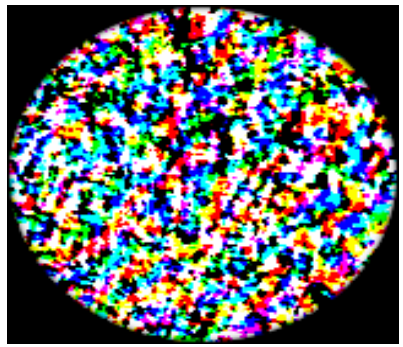


Figure 119: truck class

# Experiments

*RQ5 - Can the evidential uncertainty estimation method be used to defend against adversarial patch attacks?*

## Confusion matrix after adversarial patch attack

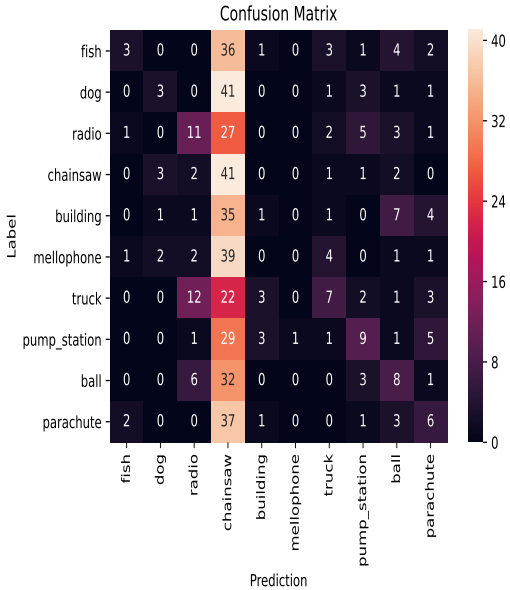



Figure 120: radio class

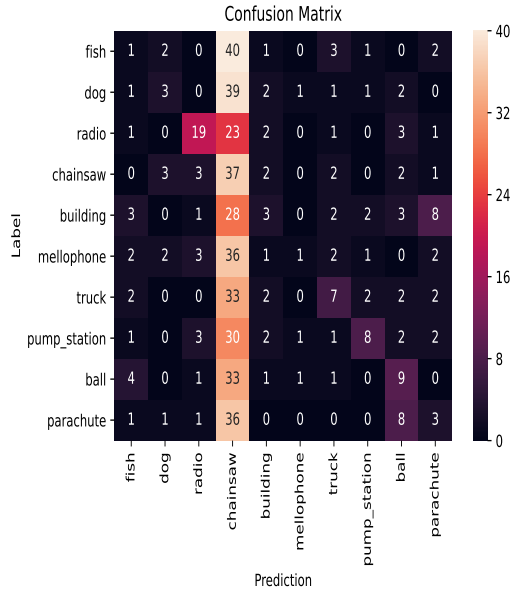



Figure 121: mellaphone class

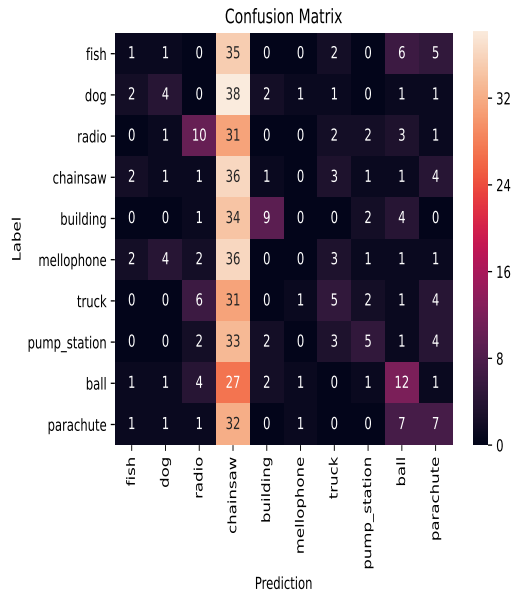



Figure 122: truck class

# Experiments

*RQ5 - Can the evidential uncertainty estimation method be used to defend against adversarial patch attacks?*

## Confusion matrix after adversarial patch attack

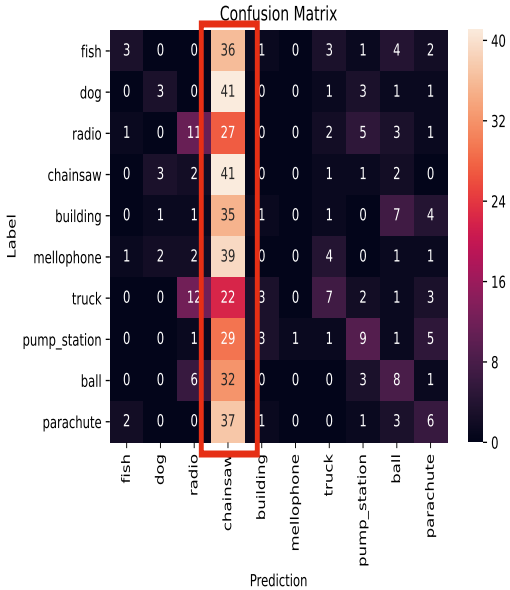



Figure 123: radio class

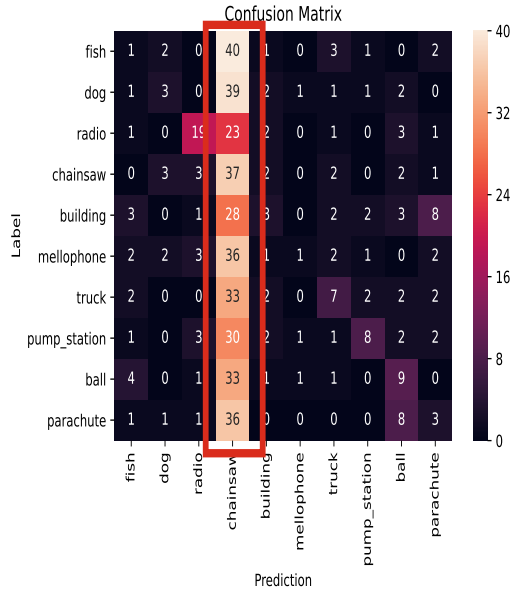



Figure 124: mellaphone class

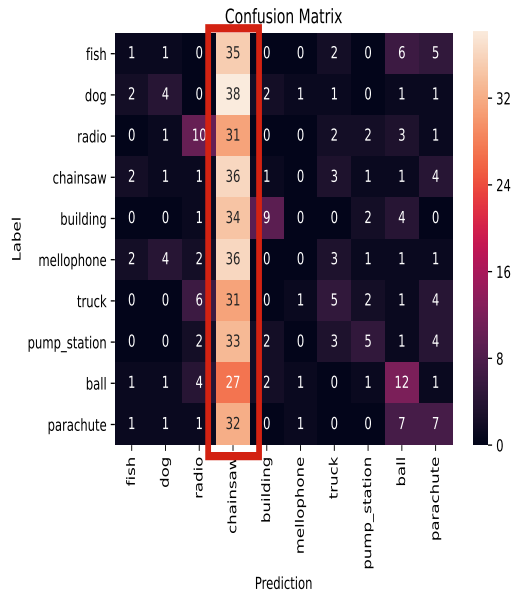
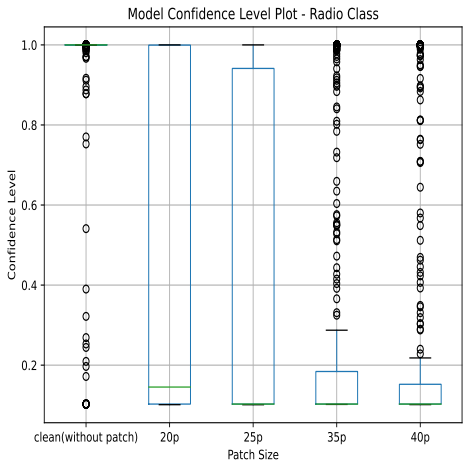


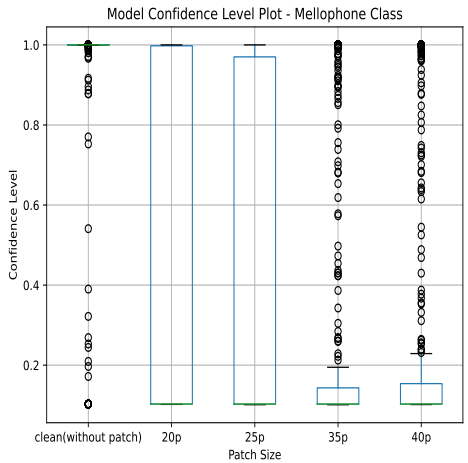

Figure 125: truck class

# Experiments

*RQ5 - Can the evidential uncertainty estimation method be used to defend against adversarial patch attacks?*

Confidence level of final predictions



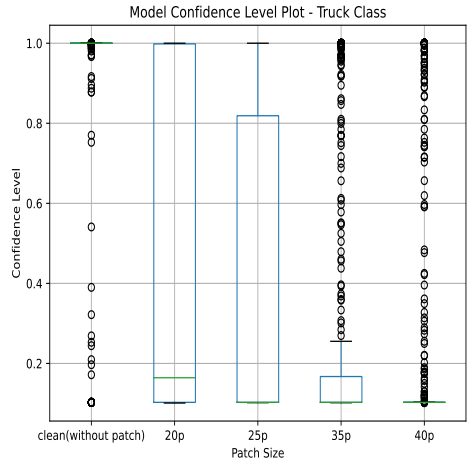




Figure 126: radio class

Figure 127: mellaphone class

Figure 128: truck class

# Experiments

*RQ5 - Can the evidential uncertainty estimation method be used to defend against adversarial patch attacks?*

Confidence level of final predictions

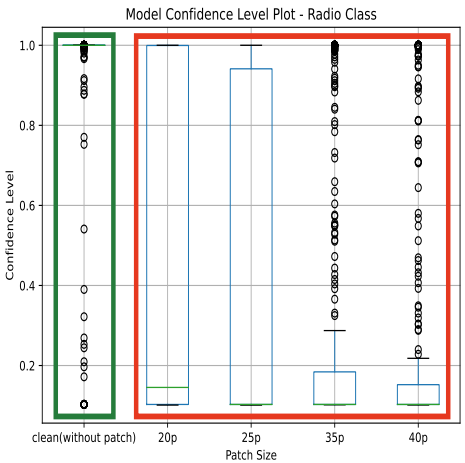



Figure 129: radio class

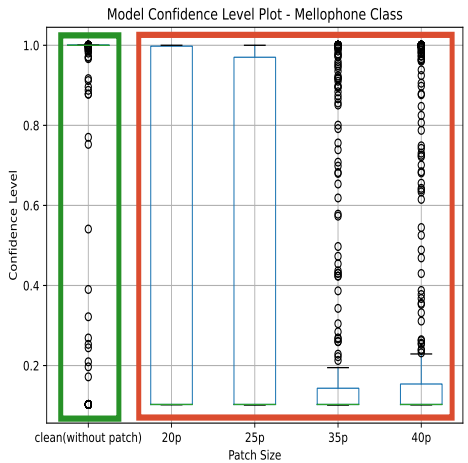



Figure 130: mellaphone class

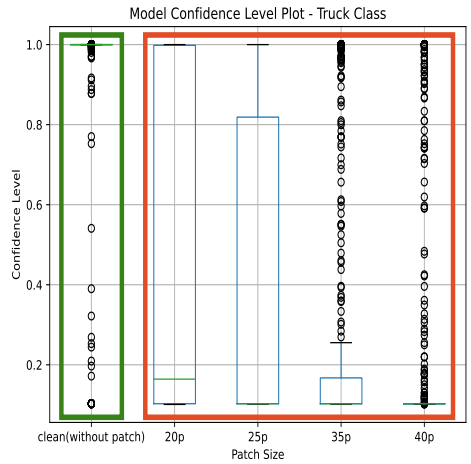



Figure 131: truck class

# Table of Contents

# Conclusion

## *Evaluation summary*

| Experiment | Model | Datasets | | |
|---|---|---|---|---|
| | | **Beans** | **Imagenette** | **RoboCup@Work** |
| Adversarial training defense | MobileNetV2 | Not effective | Not effective | Not effective |
| | ResNet50 | Not effective | Not effective | Not effective |
| | VGG16 | - | Not effective | Not effective |
| Abstention class defense | MobileNetV2 | Effective | Effective | Effective |
| | ResNet50 | Effective | Effective | Effective |
| | VGG16 | Not effective | Not effective | Effective |
| Evidential uncertainty estimation defense | MobileNetV2 | Effective | Effective | - |
| | ResNet50 | Effective | - | - |

Table 1: List of all defense experiments performed against adversarial patch attack

# Conclusion

## *Contributions*

- Literature on defense mechanisms against adversarial patch attacks

# Conclusion

*Contributions*

- Literature on defense mechanisms against adversarial patch attacks
- Targeted adversarial patch attack

# Conclusion

*Contributions*

- Literature on defense mechanisms against adversarial patch attacks
- Targeted adversarial patch attack
- Adversarial training defense

# Conclusion

*Contributions*

- Literature on defense mechanisms against adversarial patch attacks
- Targeted adversarial patch attack
- Adversarial training defense
- Abstention class defense

# Conclusion

*Contributions*

- Literature on defense mechanisms against adversarial patch attacks
- Targeted adversarial patch attack
- Adversarial training defense
- Abstention class defense
- Evidential uncertainty estimation defense

# Conclusion

## *Future Work*

- Transferability of adversarial training

# Conclusion

*Future Work*

- Transferability of adversarial training
- Extension of abstention class defense
  - type of data
  - number of data

# Conclusion

*Future Work*

- Transferability of adversarial training
- Extension of abstention class defense
  - type of data
  - number of data
- Evidential uncertainty estimation defense can be extended

# Table of Contents

# References


- Naveed Akhtar and Ajmal Mian. “Threat of adversarial attacks on deep learning in computer vision: A survey”. In: **Ieee Access** 6 (2018), pp. 14410–14430.
- Anish Athalye, Nicholas Carlini, and David Wagner. “Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples”. In: **International conference on machine learning**. PMLR. 2018, pp. 274–283.
- Tao Bai et al. “Recent Advances in Adversarial Training for Adversarial Robustness”. In: **arXiv preprint arXiv:2102.01356** (2021).
- Tom B Brown et al. “Adversarial patch”. In: **arXiv preprint arXiv:1712.09665** (2017).
- Nicholas Carlini and David Wagner. “Audio adversarial examples: Targeted attacks on speech-to-text”. In: **2018**

**IEEE Security and Privacy Workshops (SPW)**. IEEE. 2018, pp. 1–7.


Nicholas Carlini and David Wagner. “Towards evaluating the robustness of neural networks”. In: **2017 ieee symposium on security and privacy (sp)**. IEEE. 2017, pp. 39–57.


Shang-Tse Chen et al. “Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector”. In: **Joint European Conference on Machine Learning and Knowledge Discovery in Databases**. Springer. 2018, pp. 52–68.


Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples”. In: **arXiv preprint arXiv:1412.6572** (2014).


Samuel Harding et al. “Human Decisions on Targeted and Non-Targeted Adversarial Sample.”. In: **CogSci**. 2018.


Shahar Hoory et al. “Dynamic adversarial patch for evading object detection models”. In: **arXiv preprint arXiv:2010.13070** (2020).


Jeremy Howard. **imagenette**. URL: https://github.com/fastai/imagenette/.


Anil K Jain, Debayan Deb, and Joshua J Engelsma. “Biometrics: Trust, but Verify”. In: **arXiv preprint arXiv:2105.06625** (2021).


Hyun Kwon et al. “Multi-targeted adversarial example in evasion attack on deep neural network”. In: **IEEE Access** 6 (2018), pp. 46084–46096.


Makerere AI Lab. **Bean disease dataset**. Jan. 2020. URL: https://github.com/AI-Lab-Makerere/ibean/.


Sanglee Park and Jungmin So. “On the effectiveness of adversarial training in defending against adversarial example

attacks for image classification". In: **Applied Sciences** 10.22 (2020), p. 8079.

Naman Patel et al. "Adaptive adversarial videos on roadside billboards: Dynamically modifying trajectories of autonomous vehicles". In: **2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)**. IEEE. 2019, pp. 5916–5921.

Omid Poursaeed et al. "Generative adversarial perturbations". In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. 2018, pp. 4422–4431.

Huali Ren, Teng Huang, and Hongyang Yan. "Adversarial examples: attacks and defenses in the physical world". In: **International Journal of Machine Learning and Cybernetics** (2021), pp. 1–12.

Ali Shafahi et al. "Adversarial training for free!" In: **arXiv preprint arXiv:1904.12843** (2019).


Uri Shaham, Yutaro Yamada, and Sahand Negahban. “Understanding adversarial training: Increasing local stability of neural nets through robust optimization”. In: **arXiv preprint arXiv:1511.05432** (2015).


Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. “One pixel attack for fooling deep neural networks”. In: **IEEE Transactions on Evolutionary Computation** 23.5 (2019), pp. 828–841.


Rohan Taori et al. “Targeted adversarial examples for black box audio systems”. In: **2019 IEEE Security and Privacy Workshops (SPW)**. IEEE. 2019, pp. 15–20.


Yisen Wang et al. “On the Convergence and Robustness of Adversarial Training.”. In: **ICML**. Vol. 1. 2019, p. 2.


Eric Wong, Leslie Rice, and J Zico Kolter. “Fast is better than free: Revisiting adversarial training”. In: **arXiv preprint arXiv:2001.03994** (2020).

www.alamy.com. **Background or wallpaper with colorful geometric elements with random colors**. https://bit.ly/3gn6asf. 2021.

www.colourbox.com. **Color fluidism | Abstract random colors 16**. https://bit.ly/3z1nXwp. 2021.

www.pinterest.com. **Random color mosaic tiles. Abstract background, Stock image**. https://bit.ly/3yNrkGN. 2021.