Project 8: Invisibility Cloak for Depth Deception

Adversarial Attacks on Monocular Depth Estimation

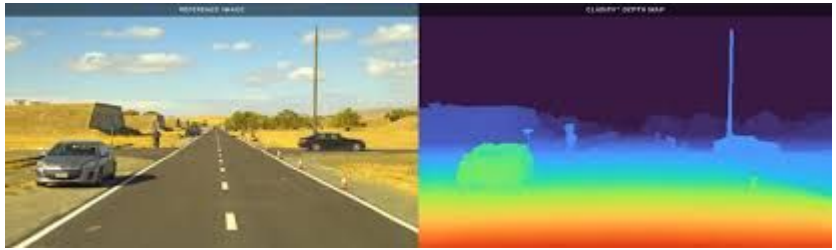
Leonardo Bisazza 1762939
Course: Computer Vision - Prof. Irene Amerini

Outline

- Problem Statement
- State of the Art
- Proposed Method
- Dataset
- Experimental Setup
- Model Evaluation (Digital & Physical)
- Conclusions

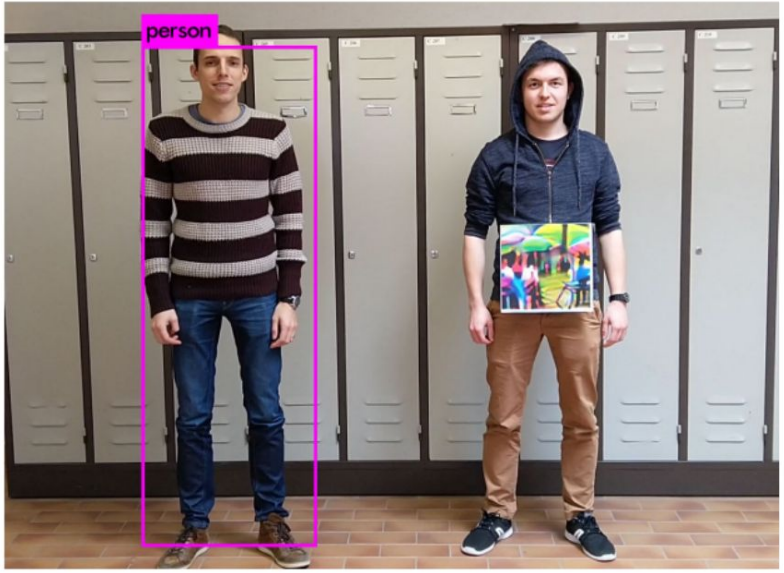
Problem Statement

- **The Context:** Monocular Depth Estimation (MDE) is critical for autonomous driving and robotics.
- **The Threat:** Can physical objects be manipulated to appear far away or disappear from the depth map?.
- **The Goal:** Adapt the "Invisibility Cloak" concept (popular in object detection) to depth regression tasks, assessing the security risks for depth-dependent applications.



State of the Art

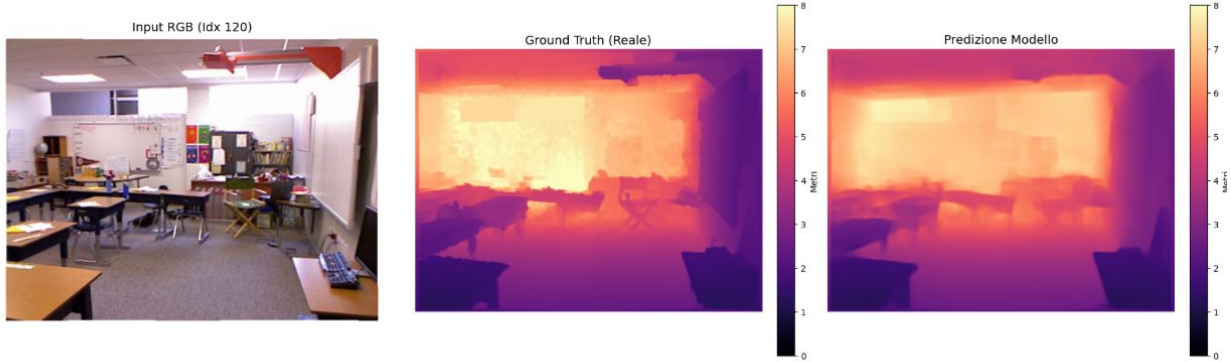
- **Adversarial Examples:** Imperceptible perturbations causing model failure (Goodfellow et al.).
- **Physical Attacks:** "Invisibility Cloak" & Adversarial Patches are proven against 2D Object Detectors (e.g., YOLO).
- **The Gap:** Vulnerability of *Depth Estimation* models is relatively unexplored compared to classification/detection.
- **Inspiration:** Thys et al. (Fooling automated surveillance cameras) .

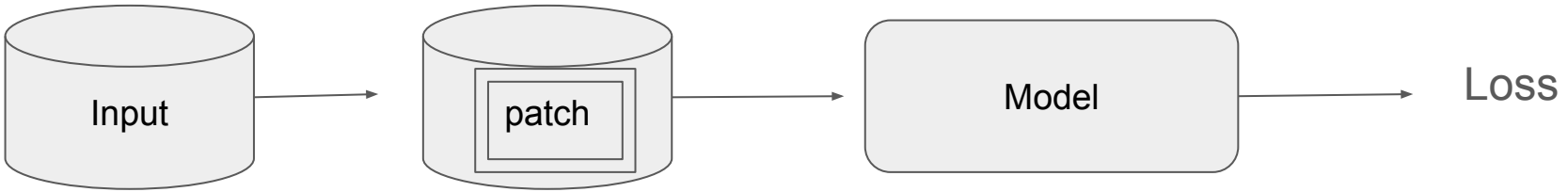


Proposed Method

1. **Baseline Training**: Implementation of *Depth Anything V2* (Encoder: ViT-Small).
2. **Attack Strategy:**
   - Objective: Maximize perceived distance (->10m).
   - Optimization: Grayscale patches (for printability) utilizing Expectation-Over-Transformation (EOT) to ensure robustness to rotation and scaling.

```
--- Campione 120 ---
GT Max:   8.358 m
Pred Max: 7.899 m
```





Dataset

Training & Digital Validation: NYU Depth V2 dataset. Used to train the baseline model and optimize the digital adversarial patch.

Physical Validation: Custom Real-World Dataset.

- Environments: Domestic scenes (Kitchen, Living Room).
- Conditions: Varying lighting (Natural, Dark), Angles (Frontal, Side), and Objects (Fridge, TV).

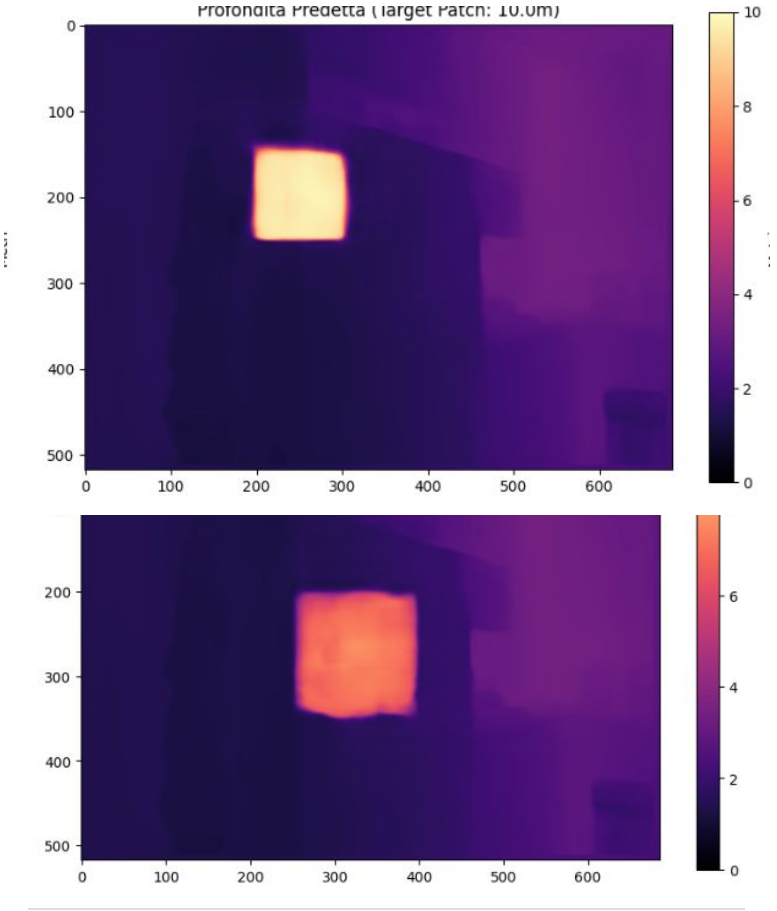
Experimental Setup - Digital Domain

Attack Config:

- **Patch Size:** 130 x 130 px.
- **Positioning**: "Sniper" Strategy (Targeting flat surfaces to avoid geometric anchors).
- **EOT:** Rotation 20°, Scaling, Jittering.

Preliminary experiments for the digital domains

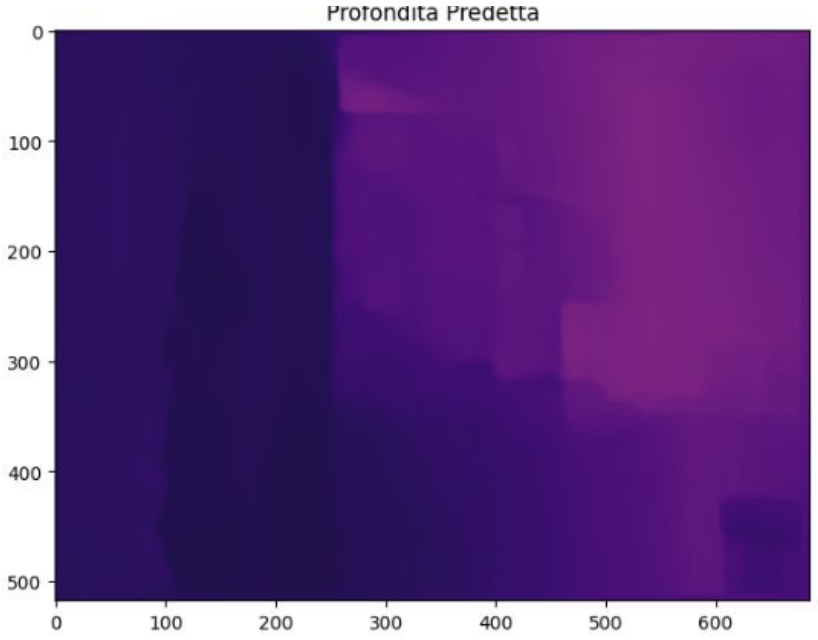
- **Static & Global Targeting:** Fixed-position patches using uniform depth targets to shift the object’s perceived distance.
- **Mobile Patch & EOT:** Optimization using Expectation-Over-Transformation with a patch that moves to find the most vulnerable geometric features.
- **Aggressive Local Jittering:** High-frequency noise combined with spatial jittering to maximize the "depth puncture" at the point of impact.



Final non printable experiment

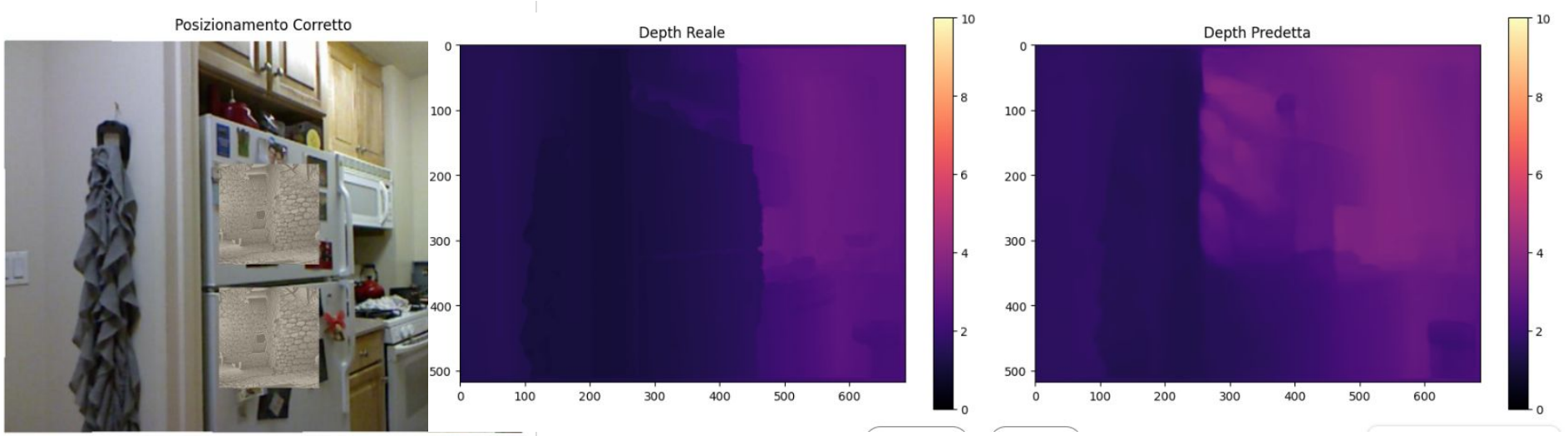
Multi-Patch Configuration: Simultaneous optimization of multiple patches to disrupt depth across the entire object surface.

Advanced EOT: Enhanced stability against extreme angles and lighting conditions using multiple coordinated patches.



Printable experiment

Multi-Patch & Binary Patterns: Ensuring physical reproduction via high-contrast distributed coverage.
Continuous Grayscale Optimization: Utilizing the full [0, 1] spectrum for complex gradients and richer adversarial textures.
Refined Sniper Strategy: Precision centroid placement to avoid "anchor features" like edges and handles.



Model Evaluation - Digital Domain

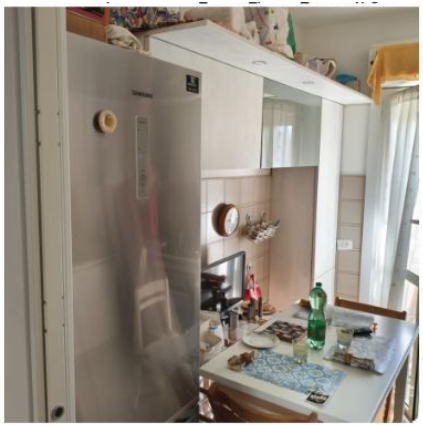
.

- **Target Achievement:** Successful transition from simple localized "depth holes" to a complete **Invisibility Cloak** effect.
- **EOT Effectiveness:** The use of EOT and spatial jittering successfully neutralized the model's ability to reconstruct depth using local textures.
- **Context Overpowering:** The attack forced the SOTA model to prioritize adversarial noise over global geometric scene consistency.
- **Digital Validation:** Proved that high-frequency noise can effectively "erase" complex volumetric objects within a 100% simulated environment.

Experimental Setup - Physical Domain

Attack Config:

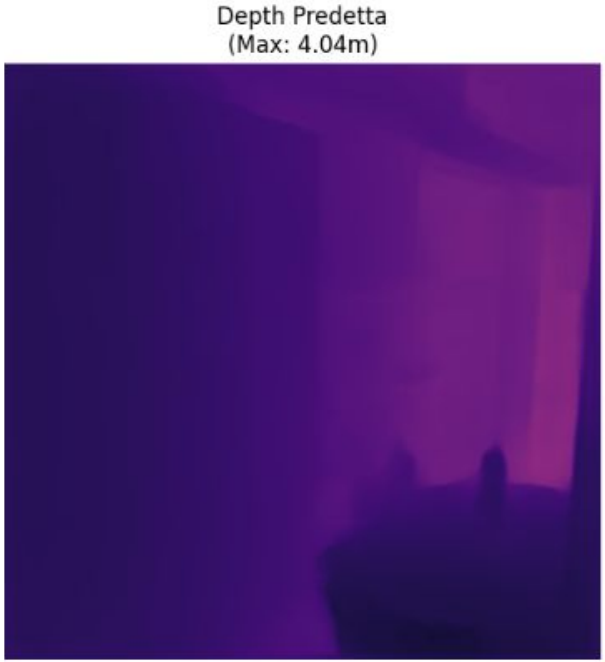
- **Patch Size:** big and small
- **Positioning**: Normal . 90° , 180°
- **Physical Setup:** Comparison between Standard Smartphone Camera (AI ISP active) vs. PRO Mode (Raw/No-Filter).



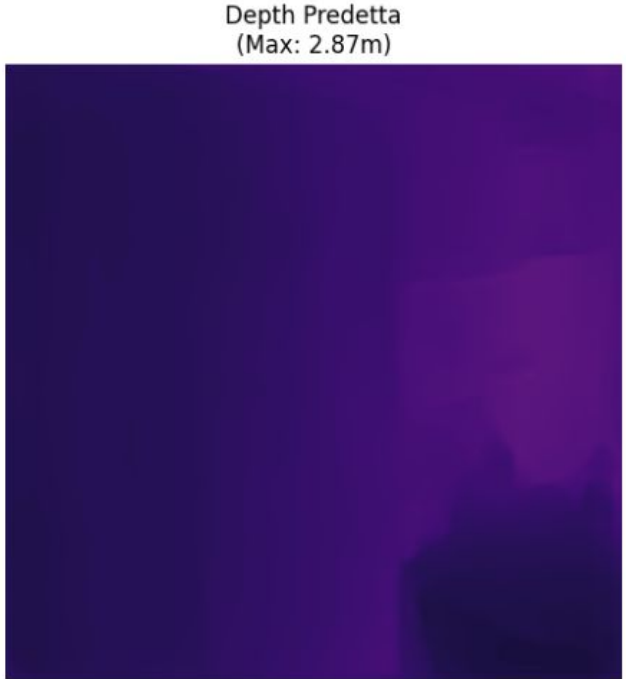

Physical Experiments - AI vs PRO picture mode

Target ->

Ai mode



Pro mode



Physical Experiments - Big vs. Small Patch

Target ->

Big Patches

Depth Predetta
(Max: 3.25m)

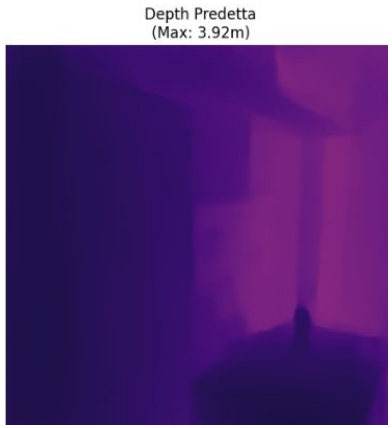
Small Patches

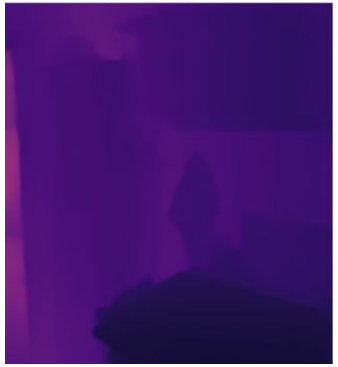
Depth Predetta
(Max: 2.87m)

Physical Experiments - Angulation

prediction

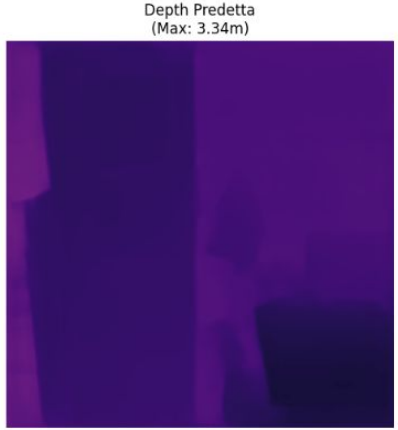
Target

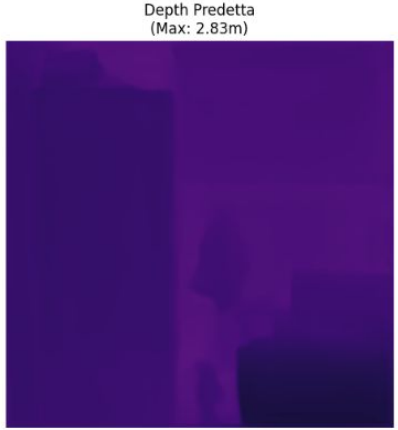




prediction

...



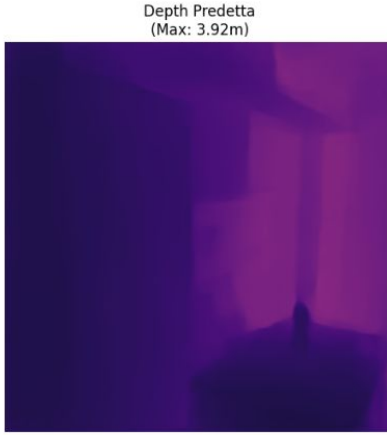


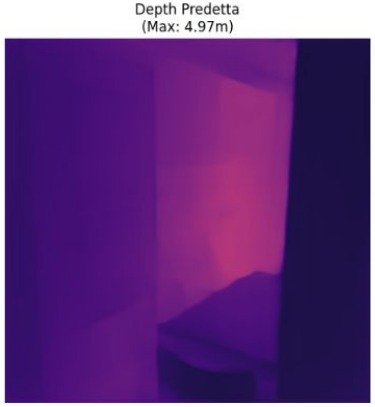
Target

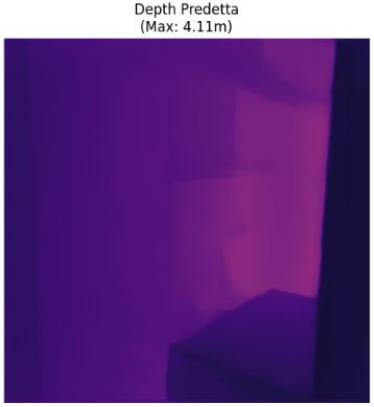
Physical Experiments - Light vs Dark

prediction

Target





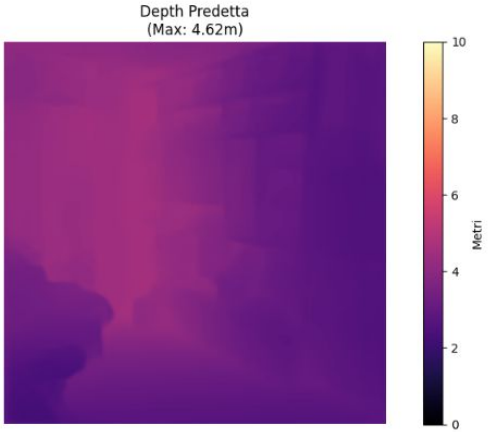


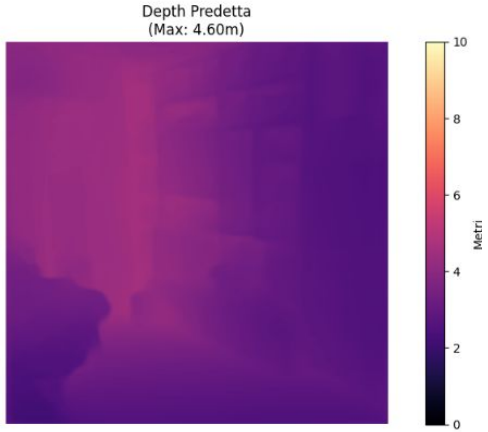
The MDE actually works better with the patch !

Physical Experiments - Tv setting

prediction

Target





Model Evaluation - Physical Domain

The Analog Barrier:

- **AI Mode:** Attack fails. Smartphone ISP (denoising/sharpening) destroys the high-frequency adversarial pattern.
- **PRO Mode:** Attack partially succeeds. Global degradation of depth estimation (scene compression), but no localized "hole".

Insight: Depth estimation networks exhibit **Contextual Robustness**. They rely on global geometry (floor, walls) rather than just local texture.

- **Effect Characterization**:
- **Lighting (Darkness):** "Backfire Effect". Patches reflect light and become *more* visible anchors, improving object detection instead of hiding it.
- **Viewing Angle:** Attack breaks at oblique angles 90°/180°. Geometry overrides texture.
- **Transferability:** Patch trained on a Fridge failed on a TV. Attacks are highly scene-specific.

Conclusions

Summary: We successfully generated a digital adversarial attack but identified significant barriers in physical transferability.

Key Findings:

1. **Context is King:** MDE models are more robust than object detectors because they leverage global 3D context.
2. **ISP Defense:** Standard camera processing acts as a natural defense against adversarial noise.

Future Work:

- Develop "Universal Adversarial Patches" trained on larger datasets.
- Investigate geometric camouflage (altering object shape) rather than just texture.

References

- Thys, S., Van Ranst, W., & Goedemé, T. (2019). Fooling automated surveillance cameras: adversarial patches to attack person detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops;
- Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018). Synthesizing robust adversarial examples. In International conference on machine learning.
- Wu, Z., Lim, S.-N., Davis, L., & Goldstein, T. (2020). Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors. arXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/1910.14667