Project 8: Invisibility Cloak for Depth Deception

Adversarial Attacks on Monocular Depth Estimation

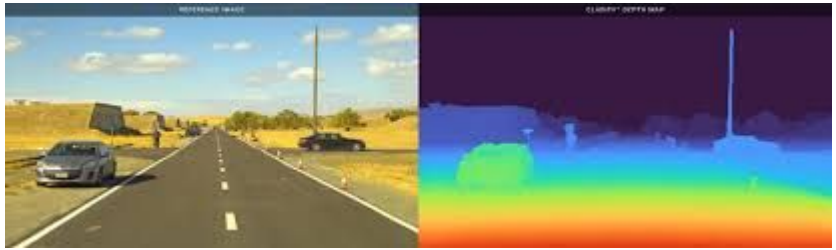
Leonardo Bisazza 1762939
Course: Computer Vision - Prof. Irene Amerini

Outline

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

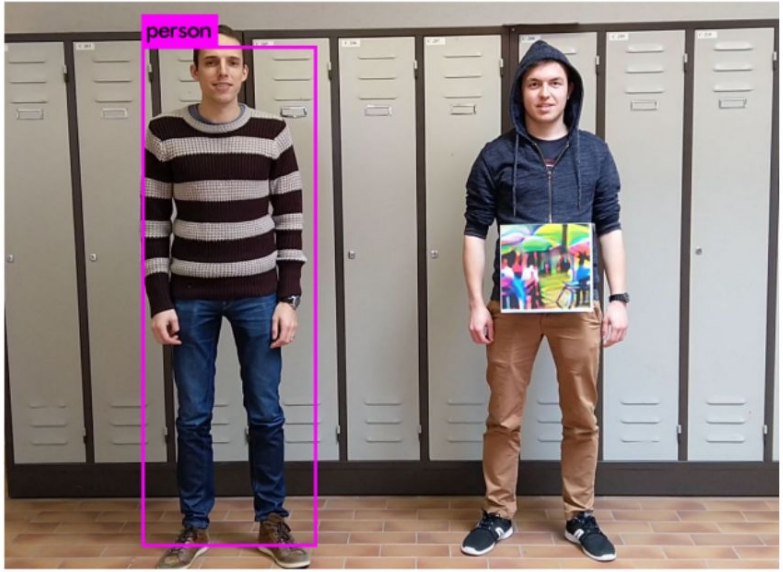
Problem Statement

- **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). +1
- **The Gap:** Vulnerability of *Depth Estimation* models is relatively unexplored compared to classification/detection.
- **Inspiration:** Thys et al. (Fooling automated surveillance cameras) .



Proposed Method

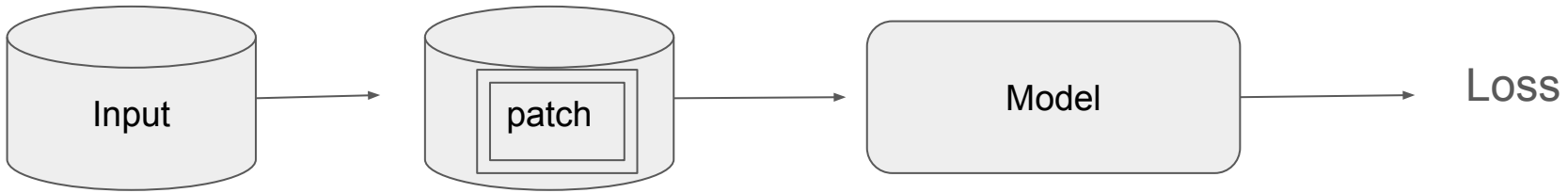
1. Baseline Training: Implementation of *Depth Anything V2* (Encoder: ViT-Small)[10].

2. Attack Strategy ("Sniper"):

- Objective: Maximize perceived distance ($D(x) \rightarrow 10m$)[11].
- Optimization: Grayscale patches (for printability) utilizing Expectation-Over-Transformation (EOT) to ensure robustness to rotation and scaling[12].

3. Pipeline: Input Image + Patch $\rightarrow$ MDE Model $\rightarrow$ MSE Loss (Target: 10m) $\rightarrow$ Gradient Update on Patch.

(Input -> Patch -> Modello -> Loss).



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.

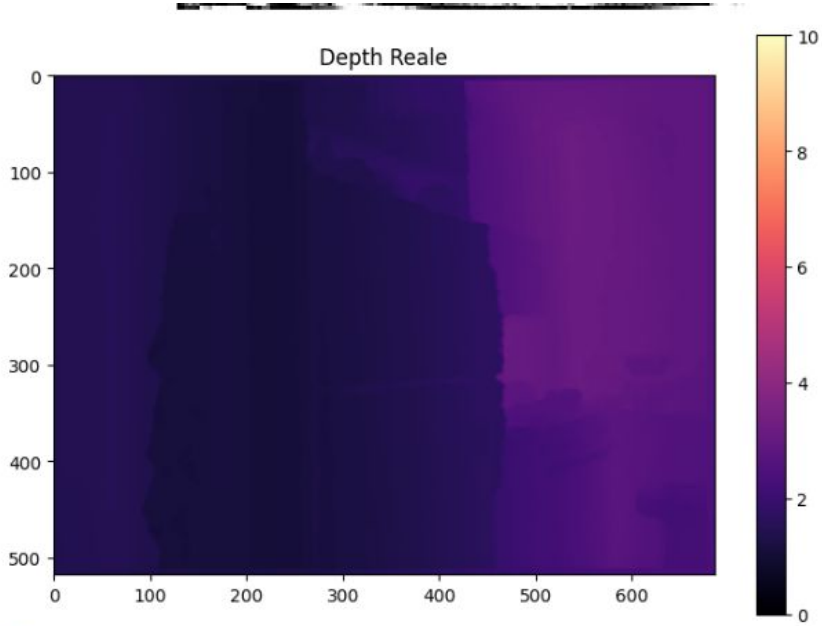
- Environments: Domestic scenes (Kitchen, Living Room).
- Conditions: Varying lighting (Natural, Dark), Angles (Frontal, Side), and Objects (Fridge, TV).

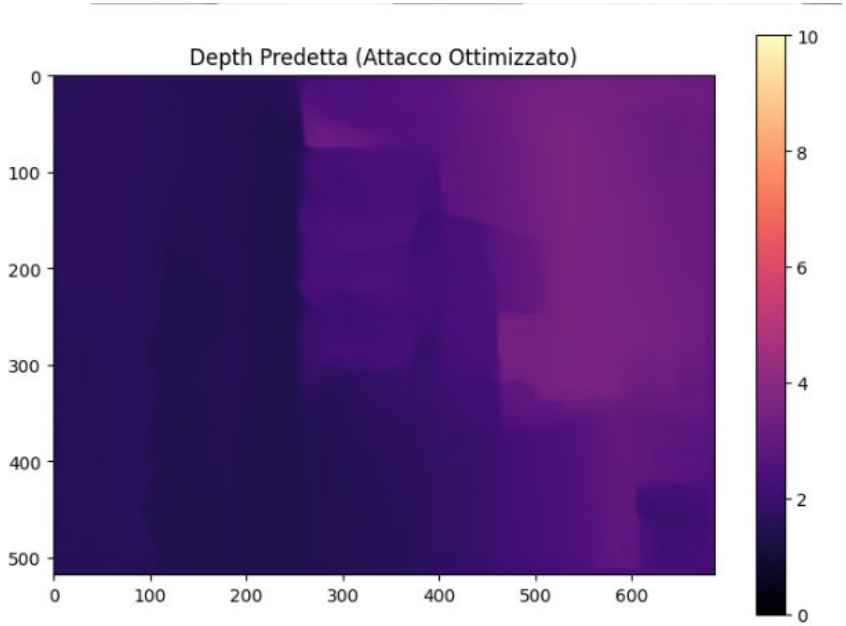
Experimental Setup

- **Hardware:** Google Colab (GPU execution).
- **Baseline Config:** L1 Loss, AdamW Optimizer, 20 Epochs.
- **Attack Config:**
- Patch Size: 130×130 px.
- Positioning: "Sniper" Strategy (Targeting flat surfaces to avoid geometric anchors).
- EOT: Rotation $\pm 20^{\circ}$, Scaling, Jittering.
- **Physical Setup:** Comparison between Standard Smartphone Camera (AI ISP active) vs. PRO Mode (Raw/No-Filter).

Model Evaluation - Digital Domain (In Silico)

- **Baseline Performance:** Accuracy $\delta < 1.25$: **0.916** (State-of-the-art performance established)[16].
- **Adversarial Success:** The digital attack successfully creates a "depth hole".
- **Impact:** Perceived depth shifts from **2m (Real)** to **>6m (Adversarial)**. The object is effectively "pushed back" in the depth map.





Model Evaluation - Physical Domain (Real World)

The Analog Barrier:

- **Standard 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.

Grafica:

- Confronto Side-by-Side:
- Sinistra: Fallimento (Standard Mode, es. `Screenshot 2026-01-08 170353.png` dove le depth ma[illegible]ono uguali).
- Destra: Successo Parziale (Pro Mode, es. `image_0d207c.png` dove la depth globale cambia).

Model Evaluation - Stress Tests & Limitations

- **Effect Characterization**[18]:
- **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 (45°). Geometry overrides texture.
- **Transferability:** Patch trained on a Fridge failed on a TV. Attacks are highly scene-specific.

Grafica:

- Usa l'immagine del test al buio (Screenshot 2025-12-03 125749.png) per mostrare come le p[illegible]iventano "fari" nella notte.

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