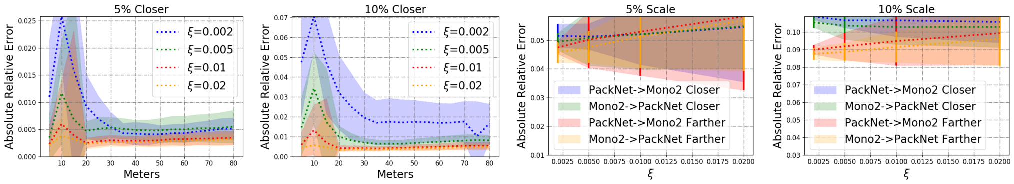

Figure 1: **Col 1, 2:** ARE binned every 5m for scaling 5 & 10% closer (5 & 10% farther also have similar plots). **Cols 3, 4:** perturbations do not transfer for Mono2 and PackNet [Guizilini et al] (also do not transfer for Mono and PackNet).

Novelty (R2). Indeed there are few papers on attacking pixelwise predictions, none of them is *targeted* in the sense of aiming to perturb specific regions, or objects and such that the output matches a target. We are the first to study *targeted* adversarial perturbations for monocular depth prediction (and dense regression tasks in general). Our findings from explainability of depth networks to properties of perturbations have not been reported by previous works. Besides the experiments in Sec. 4-6, we recommend checking Sec. 6. of Supp. Mat., where we provided additional studies on the responses of the depth networks to perturbations.

Transferability (R1, R4). We show that perturbations *do not transfer* for Monodepth and Monodepeth2 (Sec. 6). Since Monodepth and Monodepth2 share similar backbones and losses, we thought it was sufficient to demonstrate the lack of transferability and thus did not pursue further. Cols 3, 4 in Fig. 1 confirm that perturbations do not transfer for Monodepth2 and PackNet (nor for Monodepth and PackNet). We chose PackNet over VNL [Yin et al] since VNL is trained on indoors and Monodepths outdoors. We are unaware of any other *targeted* work that studies transferability.

Comparison to DAG (R2, R4). We compare extensively to DAG on 4 different attacks ($\pm 10\%$, $\pm 5\%$) each with 4 different ℓ_∞ norms (Sec. 4.1, Fig. 2, main paper). L103-115 discuss differences between the proposed method and DAG, drawbacks of DAG, and why the proposed approach is better suited for targeted dense regression attacks.

Perceptibility (R1). We apologize for the typo, $\|v(x)\|_1 = 0.362$ should be 0.0362. We cannot provide full resolution images here, but higher resolution samples are shown in Fig. 11, 15-17 of Supp. Mat. We conducted a study with 28 people where we ask if two images (image paired with itself or perturbed image, $\xi = 0.02$) are the same. Of the 40 random pairs (20 perturbed), users identified only 1 perturbed image on average, which verifies their imperceptibility.

L1/L2 norm and stability/perceptibility over time (R1). If the norm constraint is on the entire image, then perturbations may concentrate on certain regions, making it hard to guarantee imperceptibility. If it is per pixel, then the noise can concentrate on one channel, but should still be imperceptible for the same upper bounds, ξ. We don't expect flickering in the image over time since the noise is imperceptible. We will explore both directions in future work.

Metrics in meters (R1). In columns 1, 2 of Fig. 1, we bin the error every 5m. Most of the error occurs around 10m.

Missing results and selecting targets (R1). For the experiments where we had to hand select the target scenes, e.g. presets (Sec. 4.3), and manipulating instances (Sec. 5.2-5.4), we chose representative samples for qualitative results due to thousands of possible presets and manipulated instances. For preset scenes, we select 4 target scenes from the training set that are *not* similar to those in the KITTI semantic dataset (street and highway containing cars and humans) such as open roads and cluttered vegetation (Fig. 4). Here are their ARE for $\xi = 0.02$ ordered left to right according to Fig. 4: 0.0316 ± 0.0148, 0.0288 ± 0.0144, 0.0237 ± 0.086, 0.0343 ± 0.0161.

Table of conclusions (R1). Thanks, we will add a table of experiments and associated conclusions in the revised text.

Broader impact (R1). Thanks for the warning, we will revise the text to discuss these concerns.

Missing related work (R1). Thanks, we will compare against [Dijk et al.] in the revised text.

Claims (R3). We do not claim these vulnerabilities are present across all monocular depth models, but we do provide in-depth study on both supervised and self-supervised SOA methods for both indoor (VNL for NYUv2) and outdoor (Monodepth, Monodepth2 for KITTI) settings. Monodepth2 is studied extensively and we repeat the same study for VNL and again for Monodepth along with experiments on transferability. Fig. 1 includes an extra model (PackNet).

Patch-level attacks (R4). Unlike [Ranjan et al], we propose a generic *targeted* adversarial perturbations framework which can support both patch-level and global attacks. The patch is determined by the instance (Fig 6, Columns 3, 4) for our masked attacks (Sec 5.3). Moreover, we have an ℓ_∞ norm constraint so that the attacks are imperceptible.

Datasets (R4). We **do** show experiments for indoor settings (NYUv2) on VNL [Yin et al.] in Sec. 5, Supp. Mat.

Vertical flip (R4). Yes, the network (pre-trained models provided by authors) fails for vertically flipped images even without an adversary. It is true, they do not augment the image with vertical flip since it may degrade "performance" on the test set. One of the purposes in Fig. 3 is to shed light on this weakness stemming from the dataset overfitting.

White regions (R4). We provide possible explanations in L135-140 of Supp. Mat. We speculate that this is due to white being on the upper support of RGB images, which gives high activations. Hence, tweaking the intensities slightly allows perturbations to fool the network. For more details please refer to Sec. 6, Supp. Mat.