How do neural networks see depth in single images?

Tom van Dijk
Technische Universiteit Delft
J.C.vanDijk-1@tudelft.nl

Guido de Croon
Technische Universiteit Delft
G.C.H.E.deCroon@tudelft.nl

Research goal

Deep neural networks have lead to a breakthrough in monocular depth estimation. Recent work shows that neural networks can learn to predict depth from single images and the quality of these estimates is rapidly increasing. However, to the best of our knowledge no work exists that investigates *how* these networks see depth.

> "No work exists that investigates *how* these networks see depth."

Why is it important to know what these networks have learned?

1. Without this knowledge, it is difficult to guarantee correct behavior under unforseen circumstances.
2. To provide insight into training. Guidelines for the training set and data augmentation may be derived from the learned behavior.
3. To provide insight into transfer to other setups. How sensitive is the network to changes in e.g. camera pose?

In this work we take four previously published neural networks and investigate what depth cues they exploit in order to estimate the distance towards other cars in an autonomous driving setting. We use the following networks:

- **Godard et. al (2017)** (Monodepth)
- **Zhou et al. (2017)** (SfMLearner)
- **Kuznietsov et al. (2017)** (semodepth)
- **Wang et al. (2018)** (LKVOLearner)

We try to explain the behavior of the networks in terms of depth cues as observed in humans by looking for correlations between modified images of the **KITTI 2015 stereo dataset (Menze et al. 2015)** and the resulting depth predictions.

Conclusions

- All of the investigated networks use the vertical position of objects while ignoring their apparent size.
- The networks assume a fixed camera pose. Changes in camera pose are not fully accounted for by the networks and affect the estimated distance towards obstacles.
- For Monodepth the detection of objects seems to be triggered by the shadow underneath the object. The network can detect objects not present in the training set if this shadow is present.

Position vs. apparent size cues

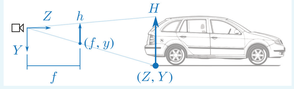

Figure 1. Geometry behind position and apparent size cues. Given the obstacle's real world size *H* and apparent size *h* in the image, the distance can be calculated using $Z = (f / h) H$. This requires the obstacle's true size *H* to be known. Alternatively, the distance can be estimated using the vertical position *y* of the object's ground contact point in the image. Given the height *Y* of the camera and the horizon level in the image y_h, the distance can be estimated through $Z = (f / (y - y_h)) Y$.

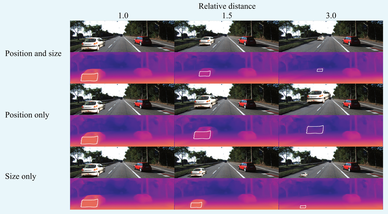

Figure 2. Test images and resulting disparity maps from Monodepth. The white car on the left is inserted at different positions and sizes. The region where the disparity is evaluated is indicated with a white outline in the disparity maps. When the position cue is present (with or without the apparent size cue), we observe a change in the estimated disparity. However, when only the apparent size cue is present the estimated disparity remains approximately constant.

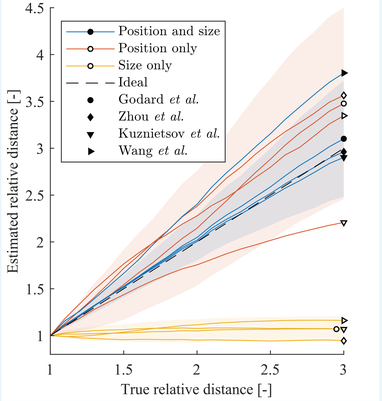

Figure 3. Influence of image position and apparent size cues on depth estimates. When both cues are present, all networks (except Wang et al.'s) correctly estimate the distance towards the objects. With only the position cue, the distance is under- or overestimated and the standard deviation increases. With only the apparent size, the networks are no longer able to estimate distance. (Shaded regions indicate ±1SD (N=1862) for the network by Godard et al.).

Depth cues

Rather than using an attribution analysis or visualization, we try to explain the behavior of the neural networks in terms of depth cues as observed in humans. Gibson (1950) and later works typically list the following depth cues as appearing in single images:

- **Position in the image**
- **Apparent size**
- Occlusion
- Texture density
- Linear perspective
- Shading and illumination
- Focus blur
- Aerial perspective ("fog")

We focus on the cues listed in **bold**; other cues are unlikely to appear because of the low image resolution (texture density, focus blur), limited depth range (aerial perspective), or because they are less relevant for absolute distance measurements (occlusion, linear perspective, shading and illumination).

> "We focus on position and apparent size; other cues are unlikely to appear in these images."

Experiment and results

To judge the influence of these cues, we measure the networks' responses when the cues are presented under conflicting conditions. We generated a dataset of modified images from the KITTI stereo dataset, where cars are inserted with one or both of the cues present (Figure 2).
The resulting distance estimates are shown in Figure 3.
All networks except Wang et al.'s correctly estimate distance on the control set where both cues are present.

> "All networks ignore the apparent size of objects in favor of their vertical position."

The performance degrades when only the position cue is present. The most surprising result is found when only the apparent size cue is present: the networks cannot observe a change in distance anymore. We find that all networks ignore the apparent size of objects in favor of their vertical position. This appears to be a general result, as this behavior is found for all four networks despite their different training regimes.

Camera pose

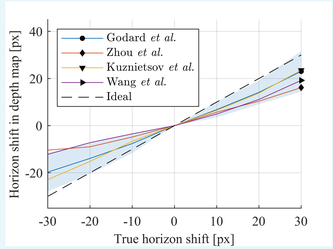

Figure 4. True and estimated changes in horizon level when pitching the camera. The change in pitch is not fully reflected in the estimated depth map.

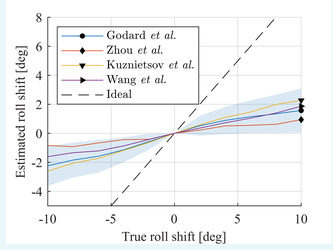

Figure 5. True and estimated changes in roll angle. The change in roll angle is not fully observed by any of the networks.

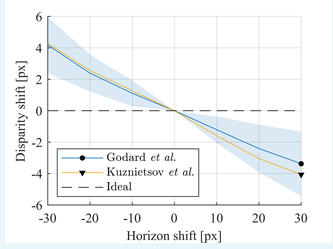

Figure 6. For the networks by Godard et al. and Kuznietsov et al., a change in pitch angle affects the estimated distance towards objects. The other networks only predict depth up to an unknown scale and are therefore not included in this experiment.

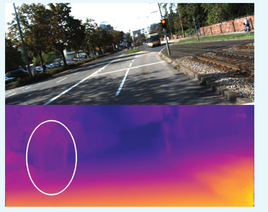

Figure 7. Qualitative example of a strong bias towards a level camera pose: the tree trunks in the disparity map appear vertical while they are clearly tilted in the original RGB image.

Experiment and results

To use the position in the image as a depth cue, the pose of the camera should be known. But is this pose learned or estimated? The pose could be inferred from the images, or learned from the training data since it is approximately constant. The answer directly affects the transfer of these networks to other camera setups.

> "The position cue requires a known camera pose. Is this pose learned or estimated?"

We generated a new dataset by cropping the KITTI images under varying pitch and roll angles:

In the resulting depth estimates, we measure the horizon level and angle of the road surface (Figure 4, 5). We find that the networks consistently underestimate the change in camera angles. This bias towards a fixed camera pose can also be seen in Figure 7.

> "The networks consistently underestimate the change in camera angles."

With the networks unable to estimate changes in camera pose, we investigate how this influences the depth estimates. We find that changes in camera pitch affect the estimated distance towards obstacles. This also suggests that the networks use the vertical position of objects rather than their distance to the visual horizon; similar results were found in experiments on humans.

> "Changes in pitch angle affect the estimated distance towards obstacles."

Obstacle recognition

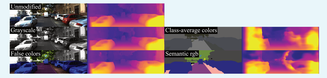

Figure 8. Removing color information or even replacing it by false colors does not significantly affect the depth estimate. Removing the texture, however, leads to significantly worse results.

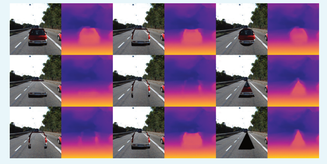

Figure 9. Not all parts of the object are required for detection; Monodepth can still recognize cars when their outline is present. Obstacles do not need a recognizable shape nor texture to be detected.

Figure 10. Objects that do not appear in the training set can be detected, but only when a dark shadow is present along their bottom edge.

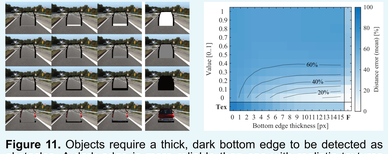

Figure 11. Objects require a thick, dark bottom edge to be detected as obstacles. A dark edge is more reliable than one with realistic textures (Tex). Completely filled shapes (F) produce the most accurate results.

Experiment and results

The use of vertical position should allow the networks to estimate distances towards arbitrary obstacles. But how does the network recognize obstacles in the first place? Figure 8 shows that the recognition depends on texture, but not on color. In Figure 9 we show that the trigger for detection is contained in the outline of the object, and that its shape or texture do not affect the detection. Figure 10 shows that arbitrary objects can be detected, but only when a shadow is present. Figure 11 shows that a thick, dark bottom edge is required for detection.

> "Obstacles require a thick, dark bottom edge to be detected."