



**What:** Monocular depth estimation method using the consistency between the stereo pair as self-supervision.

**Why:** Existing methods rely on supervised learning, and require large amounts of labeled data.

Depth estimation is a fundamental problem in machine perception. Previous monocular depth estimation methods rely on supervised learning, and applications are restricted to specific scenes (*e.g.* driving scenes), where a large amount of ground-truth depth is available. There are also some unsupervised methods, but they are not fully differentiable or can only output low-resolution depth estimations.

**How:** Train an end-to-end network that estimates disparity maps (left-to-right and right-to-left) from only the left side of the stereo image, so that the disparities to be consistent with each other (Figure 2).

Fig. 2 shows the architecture of the proposed disparity estimation module. The module only takes the left image as input, and estimates both left and right disparity maps ( $d^l$  and  $d^r$ ). Each disparity map are used to reconstruct either the right or left input image. At inference, the depth is computed from the estimated disparity by  $depth = bf/d^l$  where  $b$  is camera distance and  $f$  is focal length.

The model is trained with three loss functions. The first one is the similarity of reconstructed images is evaluated by a combination of SSIM and L1 loss (Eq. 2). The second one is the smoothness of disparity maps, which is the absolute value of gradients weighted by a term to avoid erasing valid edges (Eq. 3). The third one is the consistency loss, which is defined by L1 norm between the estimated left and right disparity maps (Eq. 4).

Fig 3 shows the whole architecture of the proposed model. The model is an encoder-decoder network. The decoder includes four disparity estimation modules ex-

plained above, each of which output disparity maps with different resolutions.

The proposed model estimates both the left and the right disparity maps. We can instead estimate only one side of the disparity map (Fig. 3 No LR), but the result suffers from artifacts in highly textured regions, as depicted in Fig. 5. The consistency loss between left and right views is effective to suppress it.

When we use the left image as input, the estimated disparity map has occlusions on its left side (Supp fig 1). To deal with the occlusion, we also obtained a disparity map with the flipped input image, and flip it back again. The flipped one has occlusions on its right side. The original and flipped disparity maps are aligned to produce the final disparity map.

**Results:** Outperform previous methods; Inference is fast; Good generalization performance;

Trained and evaluated on the KITTI, Cityscapes, and Make3D datasets. KITTI and Cityscapes contain stereo image pairs of driving scenes. Make3D contains monocular RGB-D images, and used for evaluation only. The model was trained on Titan X for 25 hours. The inference time was 28fps/sec on average.

Tab. 1 shows the result of different variations of the proposed model on the KITTI and Cityscapes dataset. The proposed model outperforms an existing unsupervised method (Deep3D). The consistency loss and the occlusion handling contribute to the accuracy (Ours No LR and Ours pp). We expanded our method to take the stereo images as input (Ours stereo), which performed better than the monocular method (Fig. 8).

Also, we compared our model with existing supervised learning based methods. For fair comparisons, input images are cropped and downsampled in the same manner for all methods. Tab. 2 and Fig. 4 show the results. The proposed methods outperform existing methods.

To examine the generalization performance, we trained our model on Cityscapes and evaluated it on the Make3D dataset. Tab. 3 and Fig. 7 show the results. The proposed model performs favorably, though not as good as the supervised methods that are finetuned on the Make3D dataset. We also applied the proposed methods to CamVid and our own dataset. Figure 9 shows the qualitative results for that.

**Thoughts:** The distance between cameras of training data must be fixed. The application examples are limited to driving scenes.

Because the disparity depends on the relative camera position of two cameras, training data must be collected with fixed stereo cameras. This paper only shows the results from in-vehicle cameras. It is uncertain that whether the model can generalize to other scenes.