Reaction report for “Homography Loss for Monocular 3D Object Detection”

Valeriy Soltan

What I like about this paper:

The paper describes the intuition behind a new loss function that drastically improves the performance of 3D object detectors. Instead of considering objects within an image individually, this paper proposes that the model be constrained by the relative position of objects within an image. I really enjoyed reading how the authors looked at existing work and underlined the deficiencies with other approaches to solving the same problem. Additionally, I was very impressed that they developed the loss to be differentiable and an extension to existing architectures such that existing work can be augmented simply by modifying the loss function. In my opinion, it makes this particular paper a lot more transformational as the developments therein described result in a performance uplift across a wide variety of strategies. It was also really interesting to read about how the mathematical properties of a homography matrix were leveraged to create a global geometric constraint.

What I don’t like about this paper:

I wasn’t able to understand a lot of the nuance presented in the paper so instead of criticizing that which I do not understand, I’ll elaborate on some points of confusion. Is the bird’s eye view (BEV) projection just another 2D image just from above? If so, what did the authors do to leverage BEV in combination with data from the camera view to glean additional information about 3D positioning? I understand that the homography matrix encodes the projection relationship between the BEV and the image plane but how does that actually impose a global constraint on the objects contained within the image. How does uniqueness in position and the preservation of collinearity imply that this is a global relationship?

One thing that I found strange is that even though the authors make the argument that monocular cameras are cost effective, they are supervising the training of a 3D detector by using correlation between 3D and 2D data. However, wouldn’t additional hardware completely circumvent the need for this translation between dimensions? By their own admission, the authors note that the challenge in the problem stems from a single image lacking depth data. I understand if this is an intellectual exercise but the effort might be misplaced as hardware sensors like LIDAR inevitably become more accessible and cost-effective.

Future directions:

One limitation that was noted in the paper was that the authors had to make the assumption that all objects were on a flat plane, simplifying the representation of points on the BEV. They note that this doesn’t scale to a lot of different scenarios and that future work will focus on removing this limitation. Furthermore, this paper was focused on developing a more performant loss. I feel like the authors could maybe try developing a custom architecture that could maximize the performance improvements of this paradigm shift, perhaps even experimenting with complementary data from additional hardware.