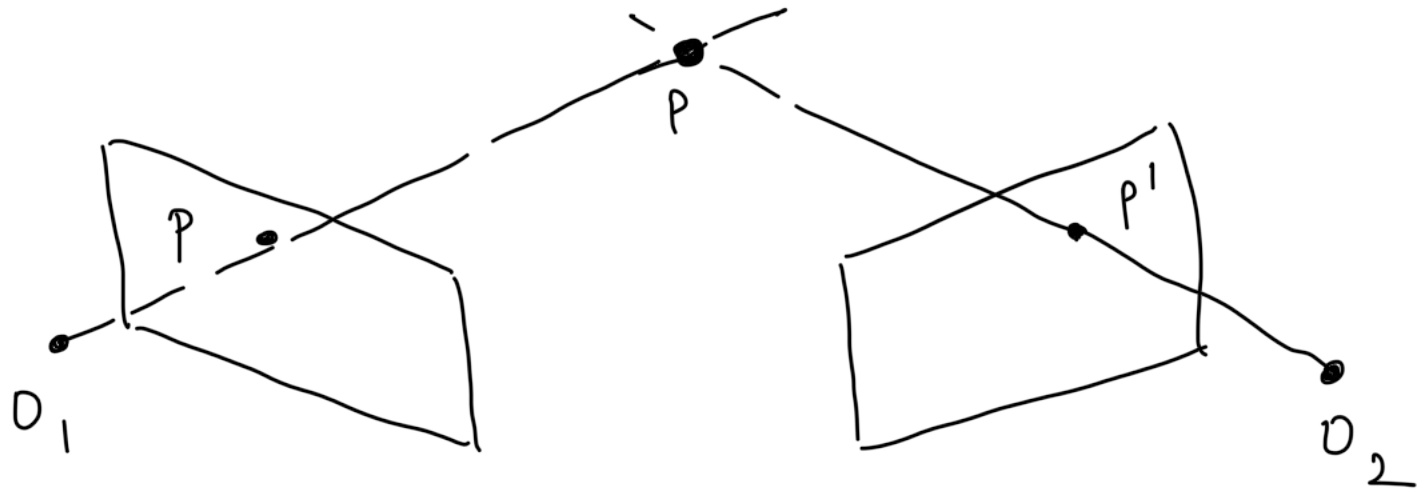
DLP - Week-11

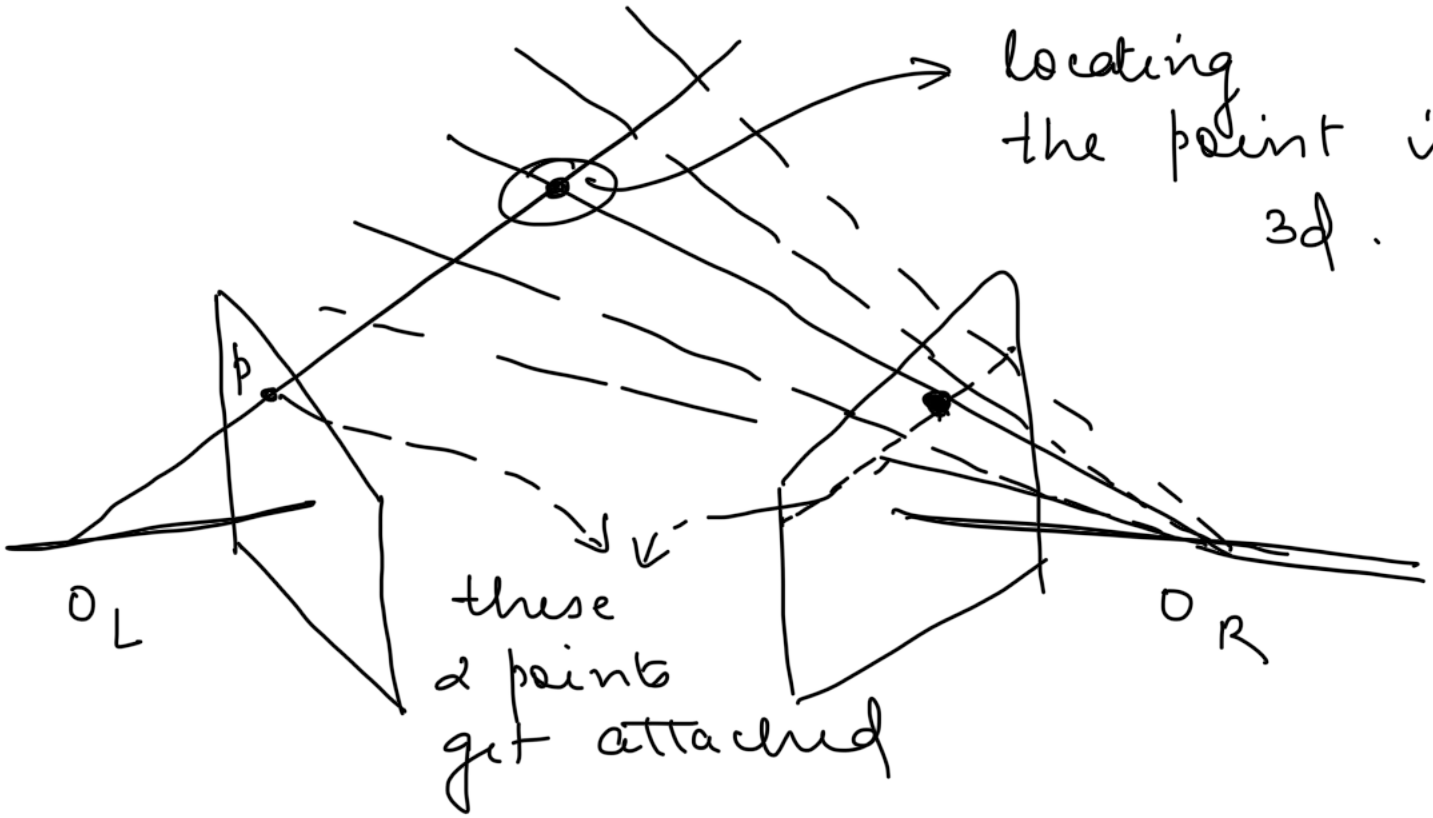
L1 Depth Estimation - Intro & Depth From Stereo

- we will be estimating the pixel-wise depth relative to the camera position
- applications of the depth estimation -
  1. autonomous navigation - indoors, roads, aerial
  2. VR/AR - virtual meeting rooms, gaming, robotic surgeries
  3. scene understanding - semantic segmentation, 3D mapping on a large scale, architectural captures
- understanding depth from stereo



(kind of a parallax error)

↳ correspondence



L2 Molecular Depth Estimation – Multi Scale Deep Network

- molecular depth estimation has few challenges —
  1. depth cues like parallax is missing this time we have only 1 image.
  2. cues available for analysis are – vanishing points, decreasing object sizes with depth.
  3. another problem, currently there is no way we can quantify these cues through a rule-based method.
- the ways to handle single image molecular depth estimation methods

1. supervised → multi scale deep network, UNet
2. unsupervised → depth from LR consistency.

- Multi Scale Deep Network —
  → network has 2 parts — coarse and fine network.
  → coarse network — predicts the overall depth from the scene.
  → refines the global predicⁿ locally (fine network)

L3 UNet

- the architecture is U shaped
- it has 2 parts — encoder downsampling & decoder upsampling network.

- encoder → uses conv layers followed by max pooling to finally reduce the spatial dimension while it continue to increase the no. of feature channels. It captures the high-level features & the relevant content from the image.

- decoder → it has the transposed conv. layers to increase the spatial dimensions of feature maps. It goes on to combine low-level features from the downsampling path with high level features to produce detailed output depth maps.
- skip connecⁿˢ → it directly connects corresponding layers in the down-sampling path to the upsampling path. It preserves the spatial info. lost during downsampling. Further, allowing for more precise reconstruc. in the output.
- loss funcⁿ — MSE → it quantifies the diff. btw. predicted depth values and ground truth values. → total no. of pixels.

$$MSE = \frac{1}{n} \sum_{i=1}^{N} (d_i - \hat{d}_i)^2$$

(d_i: true value; $\hat{d}_i$: predicted value)

- loss funcⁿ — L_depth (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^{N} |d_i - \hat{d}_i|$$

L4 Left - Right Consistency

- problem with supervised learning - is that it requires tons of training data for supervised learning. Also, since we are doing estimation, we need the image depth map pairs. (double the data).
- steps in LR consistency —
  1. use left image of the stereo pair predict the disparity
  2. true disparity isn't available during testing/training.
  3. next, transform the left to right image using the predicted disparity.
  4. true right image is present in bg during training, can be used for supervision.