Ex 6 Pen and Paper

→ Self-Supervised Optical Flow and Depth

a). ① $t=1 \rightarrow t=2$

x:

0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

y:

0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

② $t=2 \rightarrow t=3$

x:

0	0	0	1	0	0	0
0	0	0	1	0	0	0
0	0	0	1	0	0	0
0	0	0	1	0	0	0
0	0	0	1	0	0	0

y:

0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

③ $t=3 \rightarrow t=4$

x:

1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1

y:

0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

It takes me about 1 minutes to label each image, which means about 2 second for each pixel.

So it will take me about $\frac{1}{30} \times 1024 \times 1024 \times 30 \times 10$ $\approx$ 174763 hours to label a 10 second video sequence of 30 fps and resolution of 1024×1024. Thus it's obviously not feasible.

b). It is harder to perform for a shiny paint.

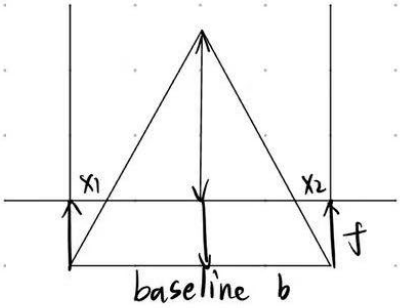
The textureless shiny parts are hard to determine the correspondence of the pixels. Also, any change of the viewpoint will change the intensity of the pixels.

c). Yes. Both the forward-backward consistency loss in the stereo model and the left-right consistency loss in the optical flow network aims to deal with the occlusion patterns and enhance the consistency.

(1). to regularize training by ensuring coherence between the forward and backward flow

(2). to remove outliers and occlusions by masking out pixels with large inconsistencies.

d).



assume the depth is z

$$d = x_1 + x_2$$

$$\frac{z-f}{b-d} = \frac{z}{b}$$

$$\Rightarrow z = \frac{fb}{d}$$

so with disparity and baseline, focal length, we can calculate the depth as above.

(2) baseline x focal length can also be estimated

by imagining an object at known depth and let the network estimate the disparity. (Directly get the pair?)

e). (1) easier to find the correspondence by methods like block matching, and therefore easier to deal with occlusions

(2) the stereo images are more consistent for they are the same scene only with a different viewpoint, while nearby frames may change a lot.

(1). avoid the challenge of handling moving objects

(2). the relative pose between a pair of stereo image is known, and need not be estimated by the algorithm during training

→ Pretext Tasks

a). (1). shortcut: consider edge continuity.

solution: add a gap between patches or jitter the patch locations

Solution: select 64x64 tiles randomly from 85x85 pixels

(2). shortcut: predict the absolute location by the abberation of color channels

solution: randomly drop some color channels or project towards gray

(3) shortcut: find by similar low-level statistics

solution: normalize the data (mean and variance)

b). The one that predicts whether vertically flipped will perform better. For the consistency between a picture and its horizontally flipped version can be high in some way and therefore can not learn the general representations.

solution: shortcuts for horizental flip
judge the 'canonical pose' in vertical flip

c). Both of them will be poor but I think vertically flipped network will still perform better.

→ Constractive Learning

a). f_i^n refers to the nth scalar of vector f_i, then:

(1). L_1 Distance:

$$D_{L_1} = \sum_{i=1}^{n} \| f_1^i - f_2^i \|$$

(2) L_2 Distance:

$$D_{L_2} = \sqrt{\sum_{i=1}^{n} (f_1^i - f_2^i)^2}$$

(3) cosine similarity:

$$D_{cos} = \frac{f_1^T \cdot f_2}{\sqrt{\sum_i^n f_1^{i\,2}} + \sqrt{\sum_i^n f_2^{i\,2}}}$$

L_2 Distance > L_1 Distance > Cosine Similarity

b). Without the negative examples, the network will overfit for the loss gets the minimum when $s(f(x), f^{+}(x))$ gets the maximum

→ Input Optimization

a). Black-box attack doesn't know the property (parameters) of the networks while white-box attack does. So black-box attack is harder.

b).

$$\begin{bmatrix} F_1^2(a)+F_1^2(b)+F_1^2(c)+F_1^2(d) & F_1(a)\cdot F_2(a)+F_1(b)\cdot F_2(b)+F_1(c)\cdot F_2(c)+F_1(d)F_2(d) \\ F_2(a)\cdot F_1(a)+F_2(b)F_1(b)+F_2(c)F_1(c)+F_2(d)\cdot F_1(d) & F_2^2(a)+F_2^2(b)+F_2^2(c)+F_2^2(d) \end{bmatrix}$$