Monocular Depth Estimation

Bjørnar Birkeland, André Håland, Torkil Solheim

The Problem

- Estimate accurate depth from a single image
- Crucial step in detection, segmentation, scene reconstruction,
  3D object recognition

Approaches

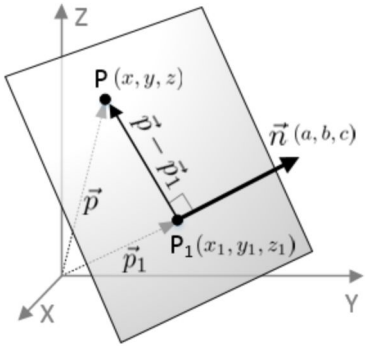
- Traditional (not single image)
  - Stereo images
  - Multiple frames from moving camera
  - Multiple static images with different lighting conditions
- Encoder - decoder, recently typically DCNN
- Feature extraction; loses resolution
- Possible solutions:
  - Multi-scale networks
  - Skip connections
  - deconvolutional networks

Multi-Scale Local Planar Guidance

State-of-the-art 2019 (late August)

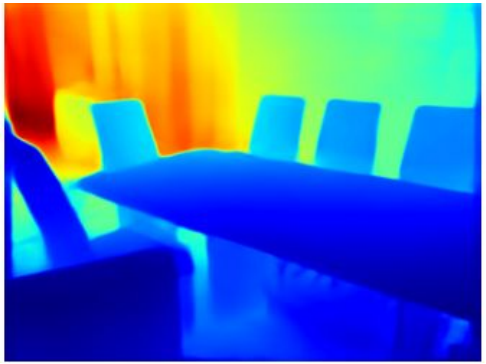
High level overview

- Input: single image
- Output: depth map



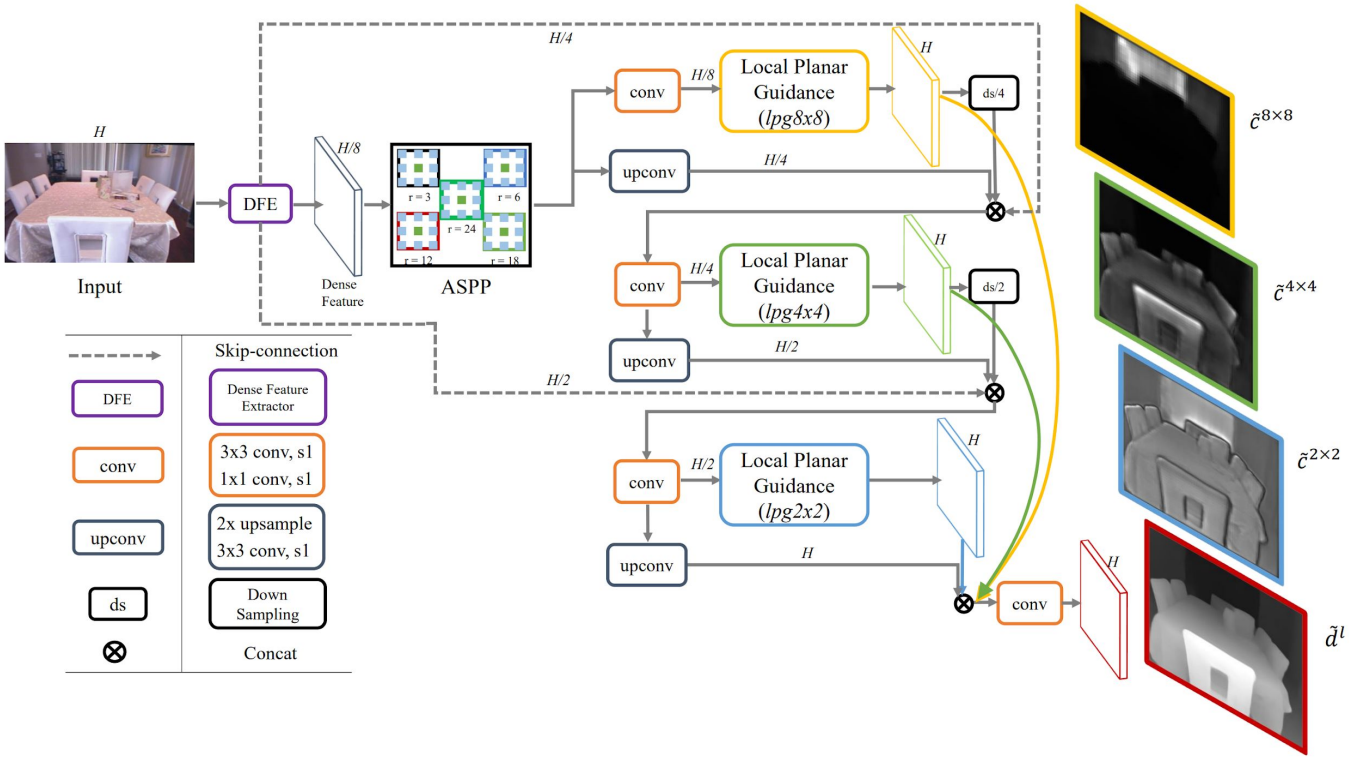
Main steps

- Feature extraction
- Segmentation
  - ASPP
- Find plane representations
  - Local Planar Guidance Layers (LPG)
  - Planes can be represented mathematically
  - Finds the depth of the plane from the images perspective



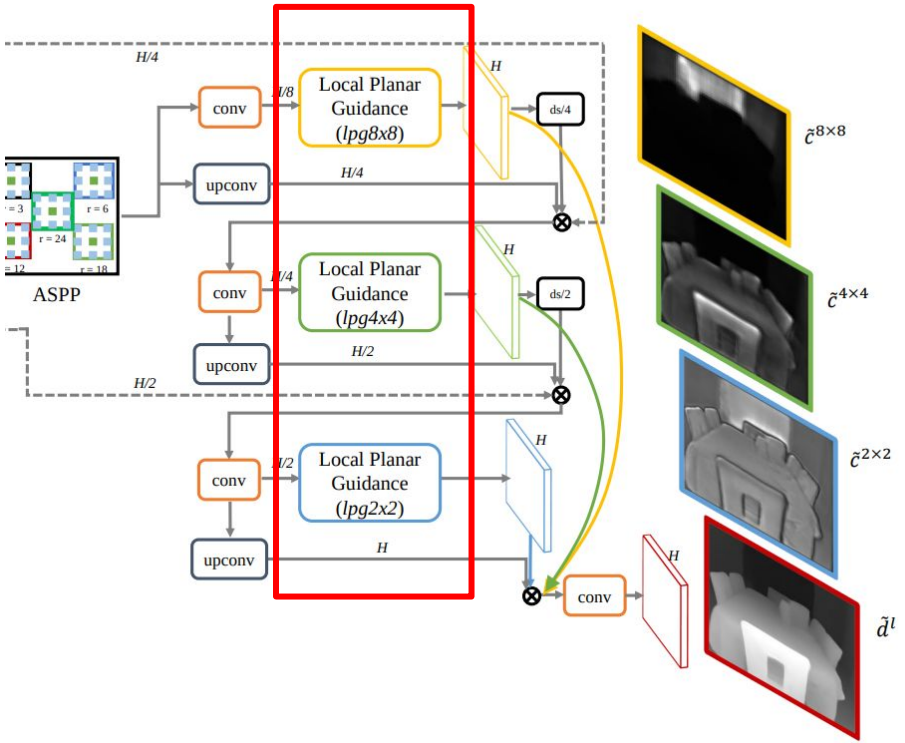
Architecture

- Encoding-decoding scheme
- Encoding
  - Dense Feature Extractor
  - Contextual information extractor (ASPP)
- Decoding
  - Local Planar Guidance trained to recognize planes
  - Three different LPGs are used and combined.



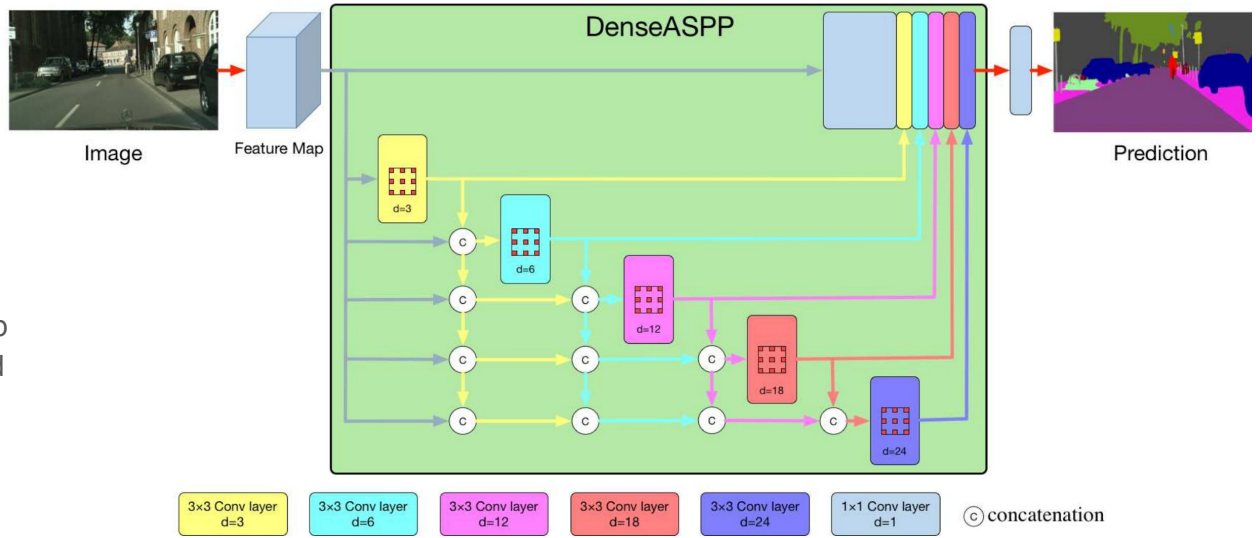
Main contribution - key idea

- Key idea
  - Define a more direct and explicit relation between internal features and the final output
- Local Planar Guidance Layers
  - New method
  - Main contribution



Contextual Information Extraction

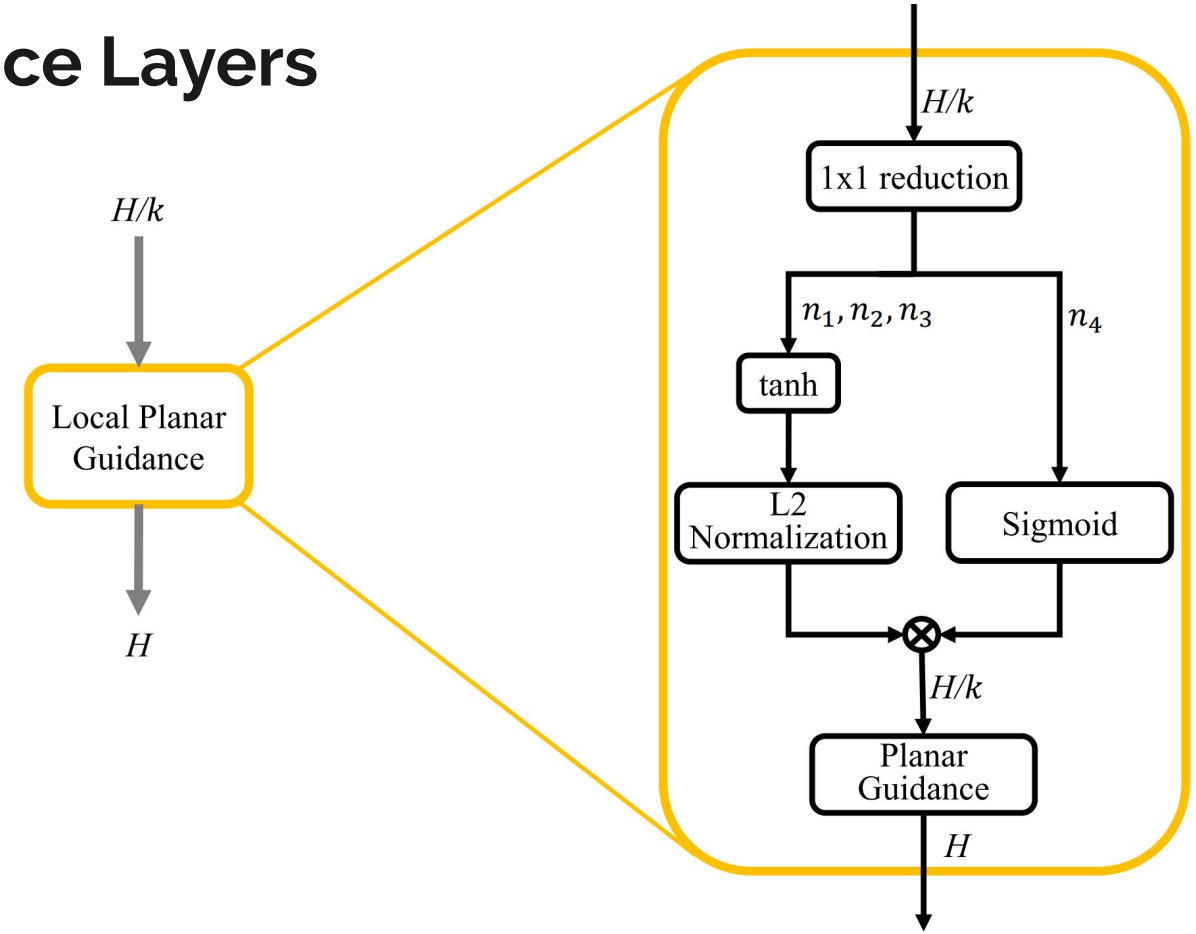
- Semantic segmentation
- Atrous convolution
  - Increase receptive field while keeping the feature map resolution unchanged.
- Atrous Spatial Pyramid Pooling
  - Concatenate feature maps with different dilation rates
  - Neurons in output feature map contain multiple receptive field sizes



[2] Yang et al., 2018, DenseASPP for Semantic Segmentation in Street Scenes

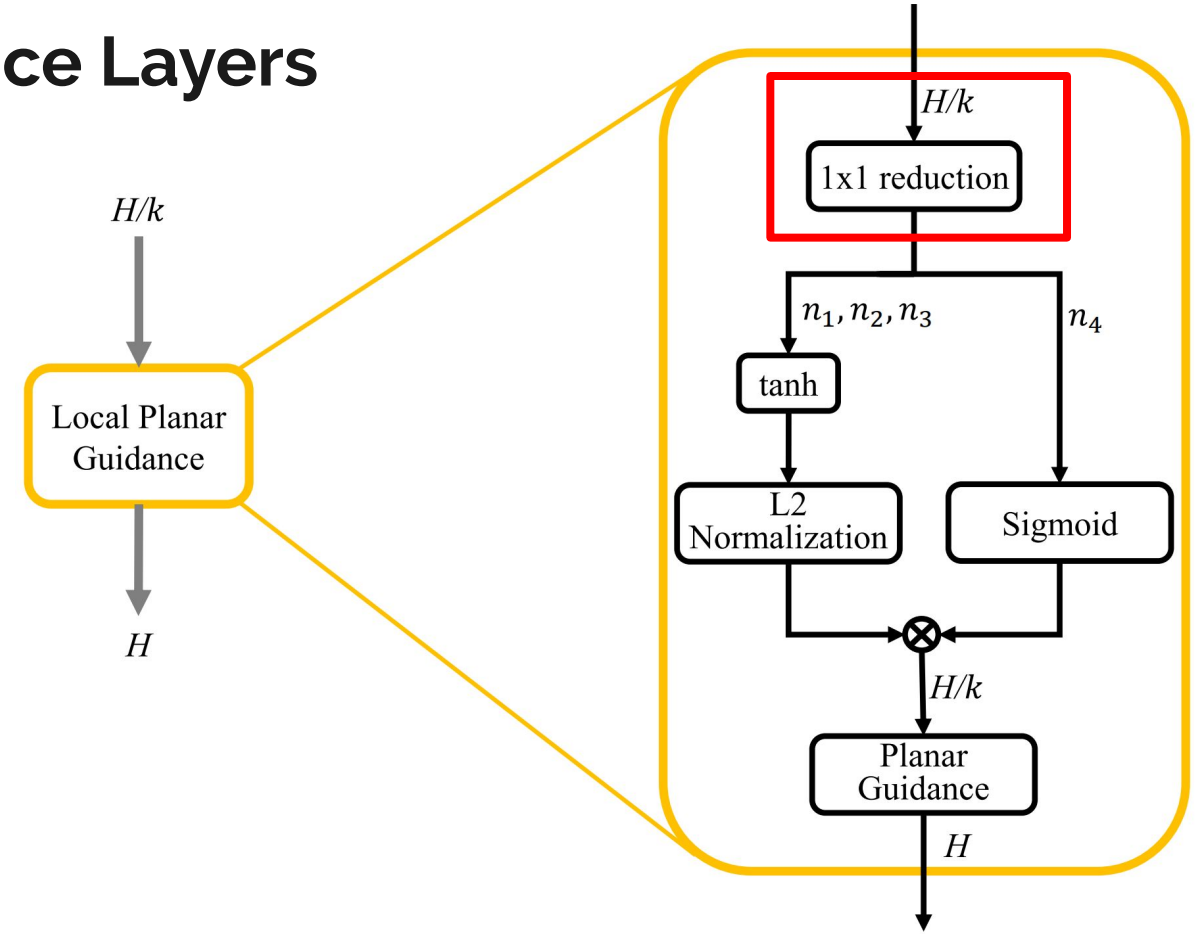
Local Planar Guidance Layers

- *"Guide features to the full resolution with the local planar assumption, and use them together to get the final depth estimation"*
- Given a feature map of size *H/k*, the layer estimate plane coefficients that fits a *k×k* patch on the full resolution H.
- Each layer only estimates local depths, and concatenated together, a global depth estimation is achieved.



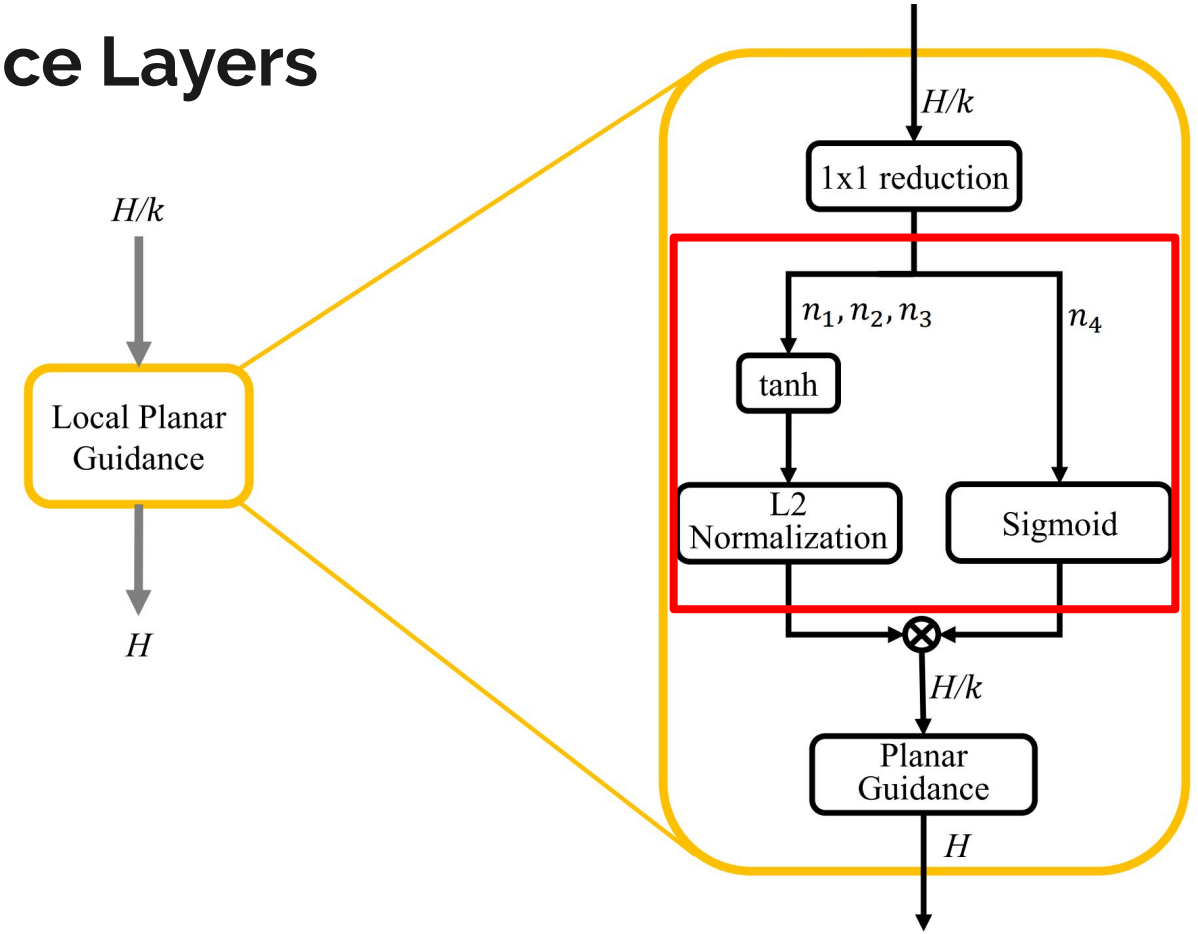
Local Planar Guidance Layers

- Reduce channels by a factor of 2 until it reaches 4
- Assuming square input, this results in $H/k \times H/k \times 4$ feature map (4D plane coefficients)



Local Planar Guidance Layers

- Coefficients are split in two ways to ensure constraints on plane coefficients

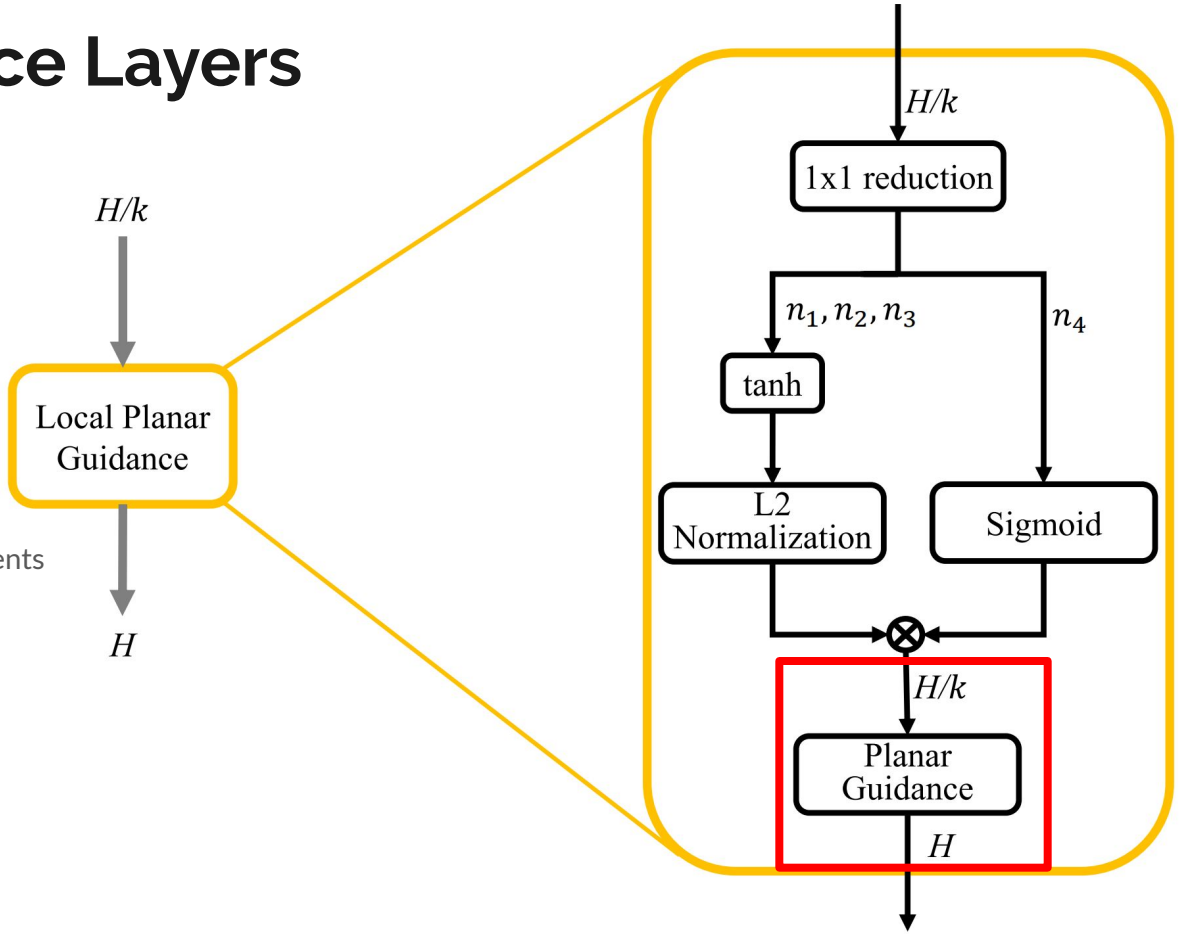


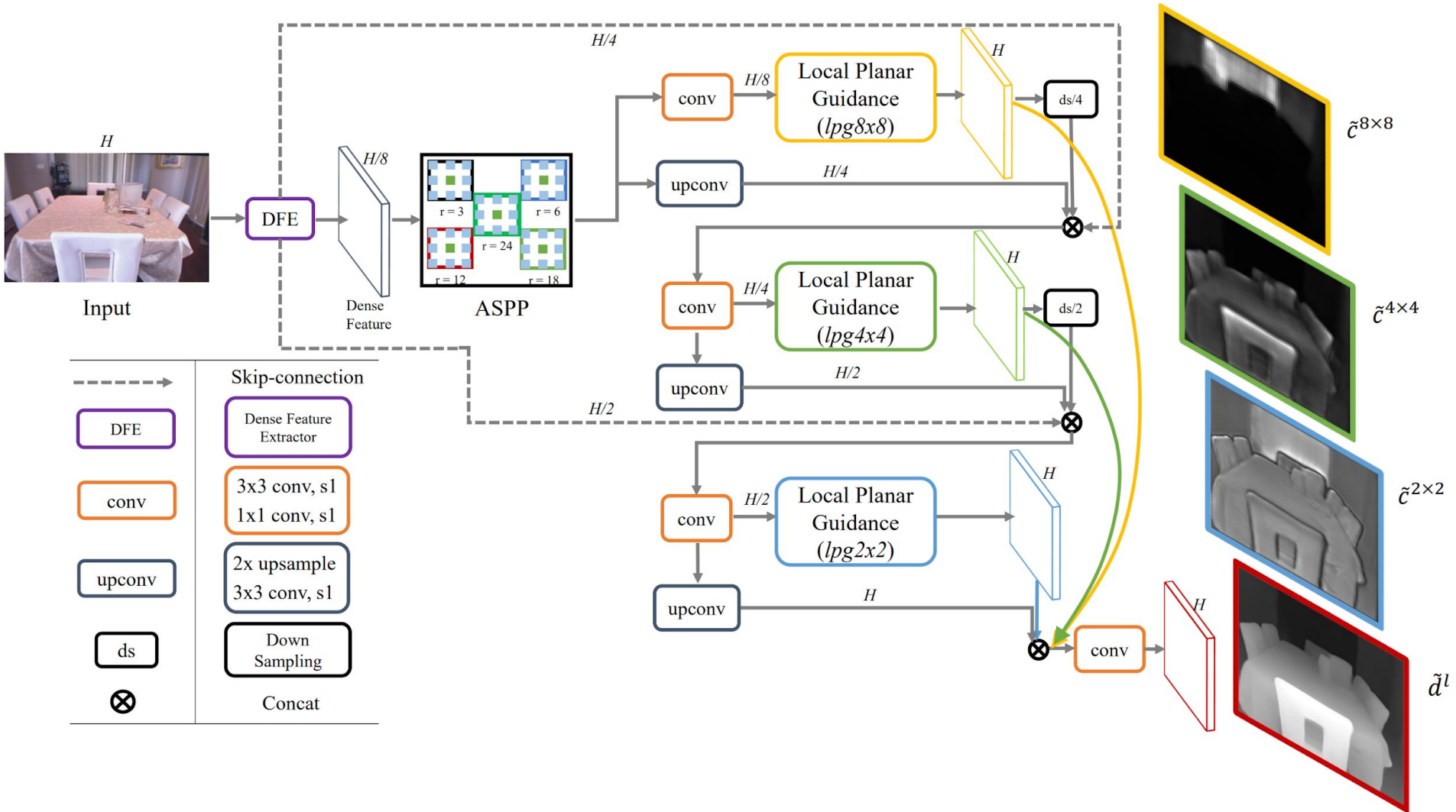
Local Planar Guidance Layers

- Local depth estimation

$$\tilde{c}_i = \frac{n_4\sqrt{u_i^2 + v_i^2 + 1}}{n_1 u_i + n_2 v_i + n_3},$$

- (n_1, n_2, n_3, n_4) is the estimated plane coefficients
- (u_i, v_i) are normalized coordinate of pixel *i*




H
Input
DFE
H/8
Dense Feature
ASPP
r = 3
r = 6
r = 24
r = 12
r = 18
H/4
conv
H/8
Local Planar Guidance (lpg8x8)
H
ds/4
upconv
H/4
conv
H/4
Local Planar Guidance (lpg4x4)
H
ds/2
upconv
H/2
H/2
conv
H/2
Local Planar Guidance (lpg2x2)
H
upconv
H
conv
H
$\tilde{c}^{8\times8}$
$\tilde{c}^{4\times4}$
$\tilde{c}^{2\times2}$
$\tilde{d}^{l}$
Skip-connection
DFE
Dense Feature Extractor
conv
3x3 conv, s1
1x1 conv, s1
upconv
2x upsample
3x3 conv, s1
ds
Down Sampling
Concat

Sources

[1] Lee et al., 2019, “From Big to Small: Multi-Scale Local Planar Guidance for Monocular Depth Estimation”
[2] Yang et al., 2018, “DenseASPP for Semantic Segmentation in Street Scenes”

Plane image:
http://www.songho.ca/math/plane/plane.html (21.10.19)