Machine Learning in Geometric Computer Vision

Part #1

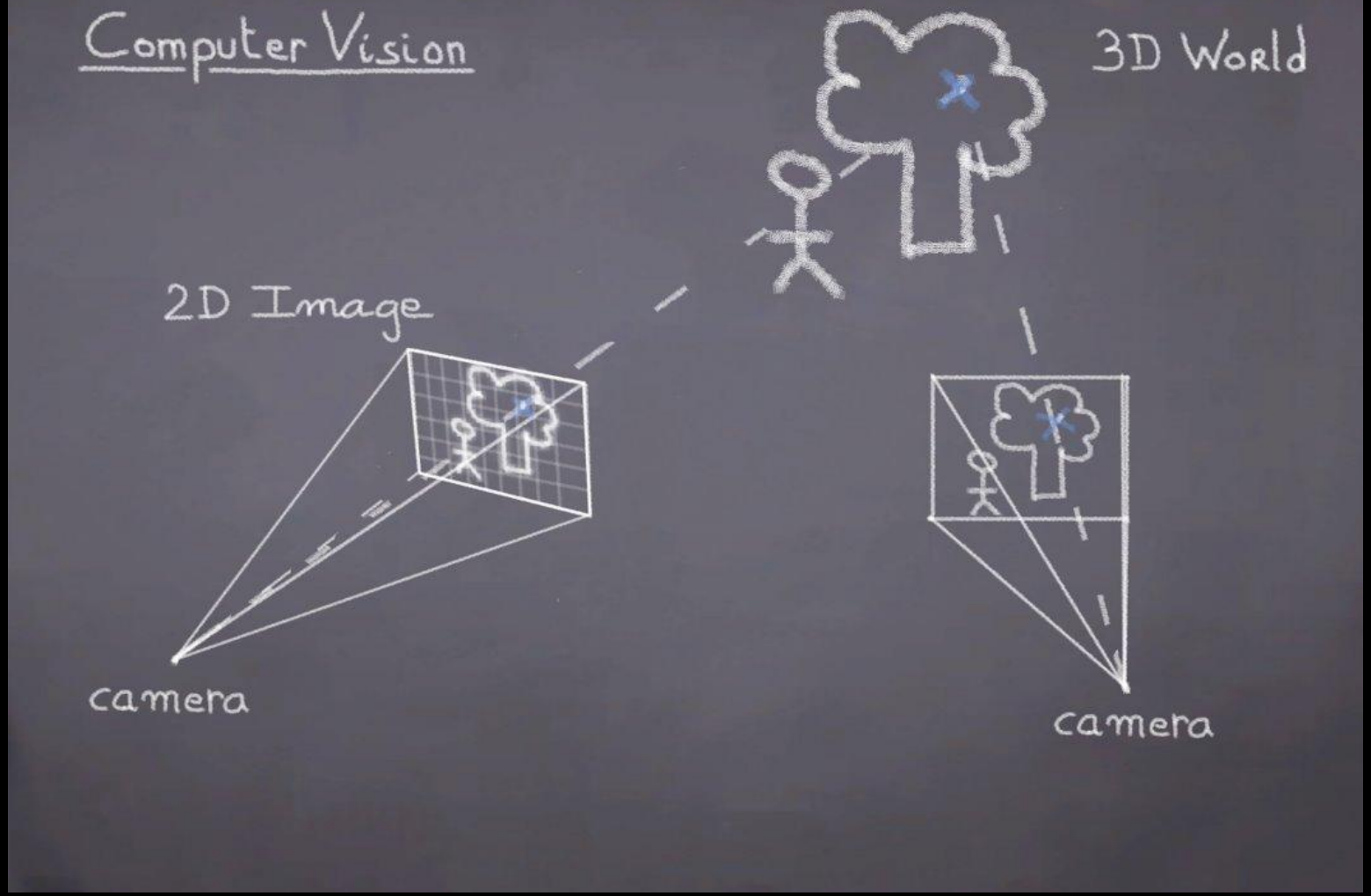
Machine Learning in Geometric Computer Vision

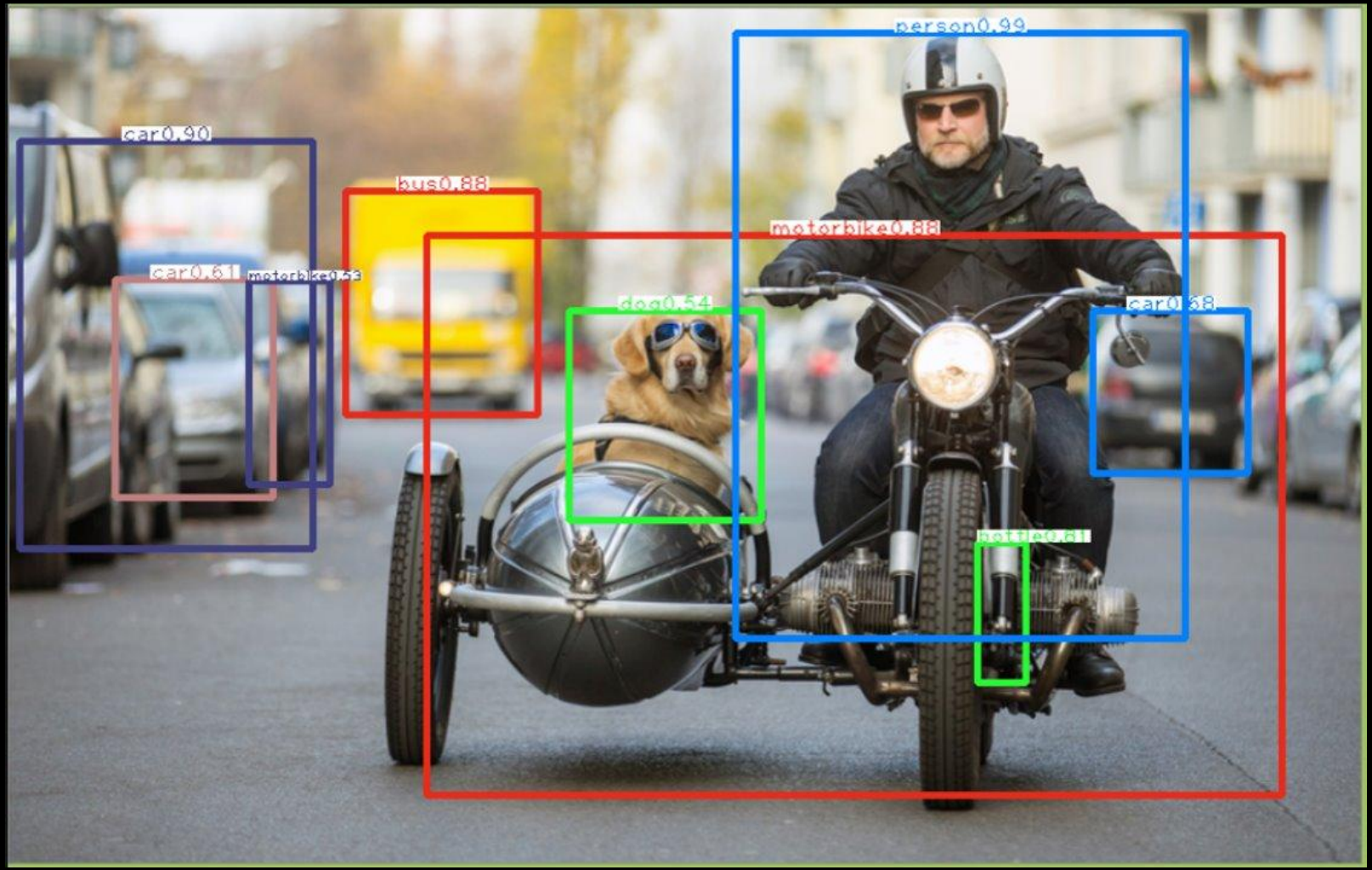
Part #1

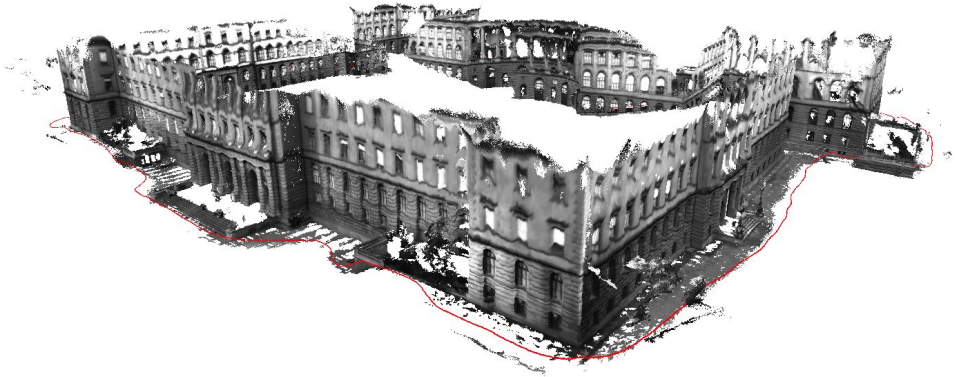
Machine Learning in Geometric Computer Vision

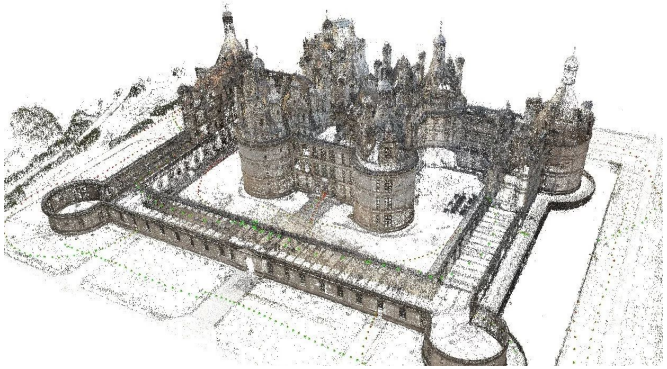
Part #1

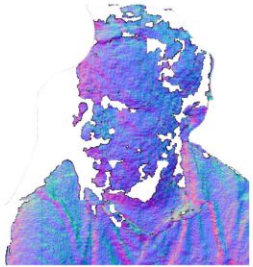
Computer Vision

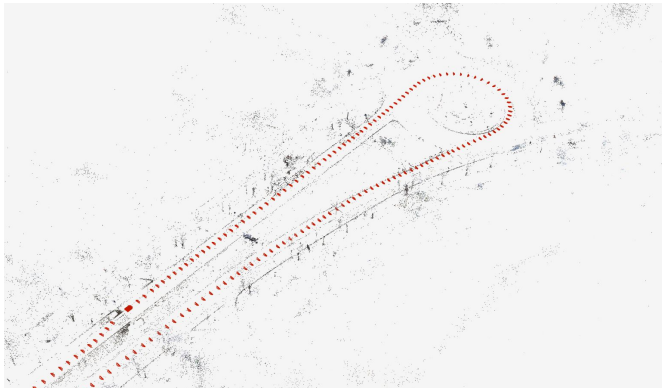


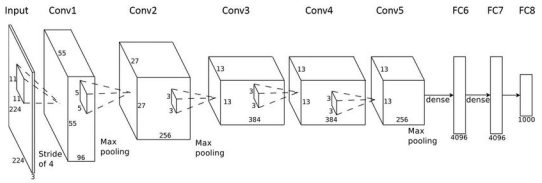

person0.99
car0.90
bus0.88
motorbike0.88
car0.61
motorbike0.56
dog0.54
car0.68







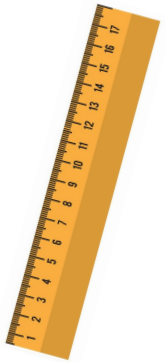



Input
Conv1
Conv2
Conv3
Conv4
Conv5
FC6
FC7
FC8
224
224
3
55
55
96
11
11
5
5
27
27
256
3
3
13
13
384
384
256
4096
4096
1000
dense
dense
Stride of 4
Max pooling
Max pooling
Max pooling

+


Nikon
Nikon

+



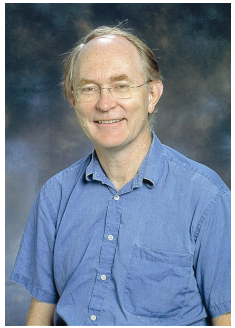
=

?




+

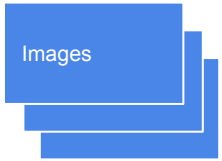


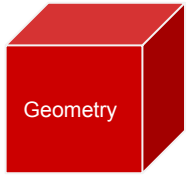
=

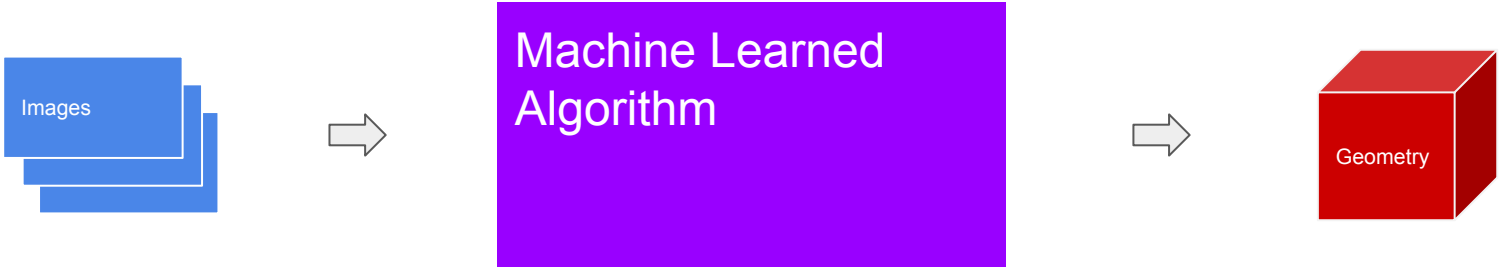
?

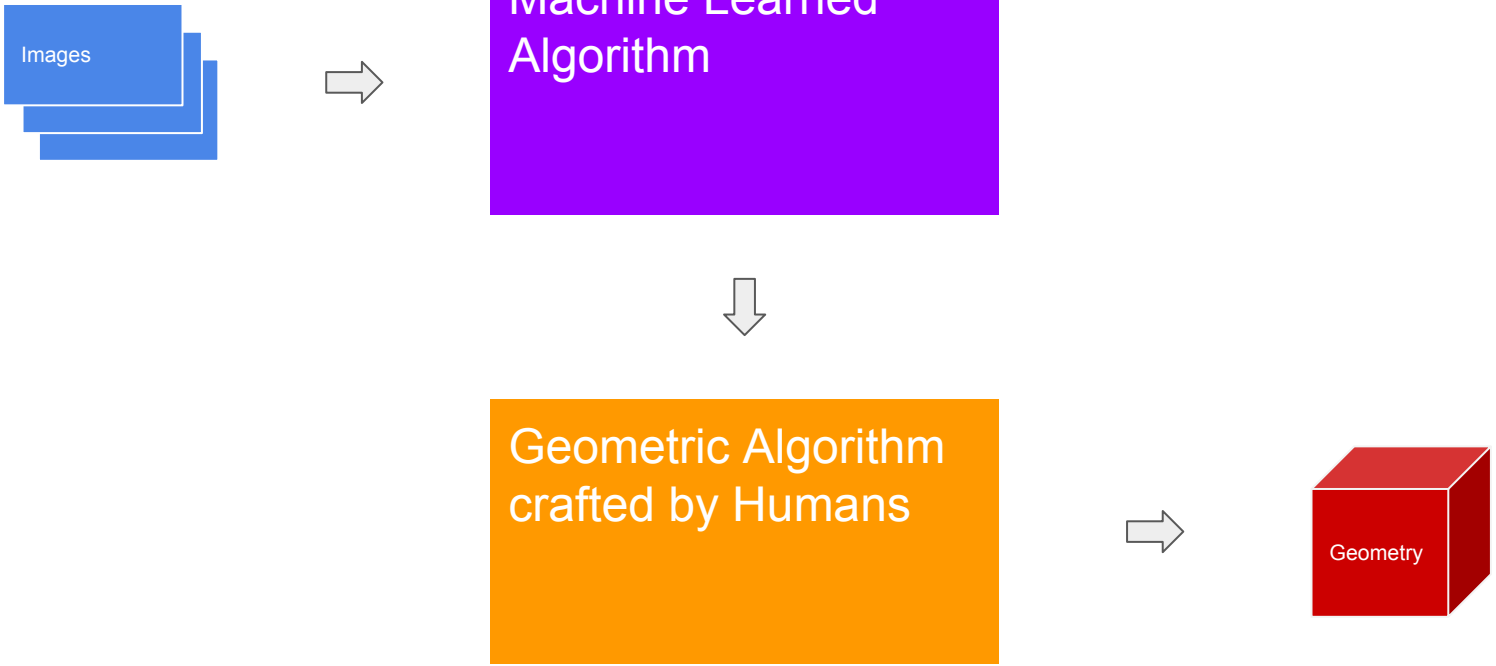
* Many great minds have been left out of this slide. Images have been resized for appearance purposes and their sizes don't correspond to any known metric of the individual's contribution to society.

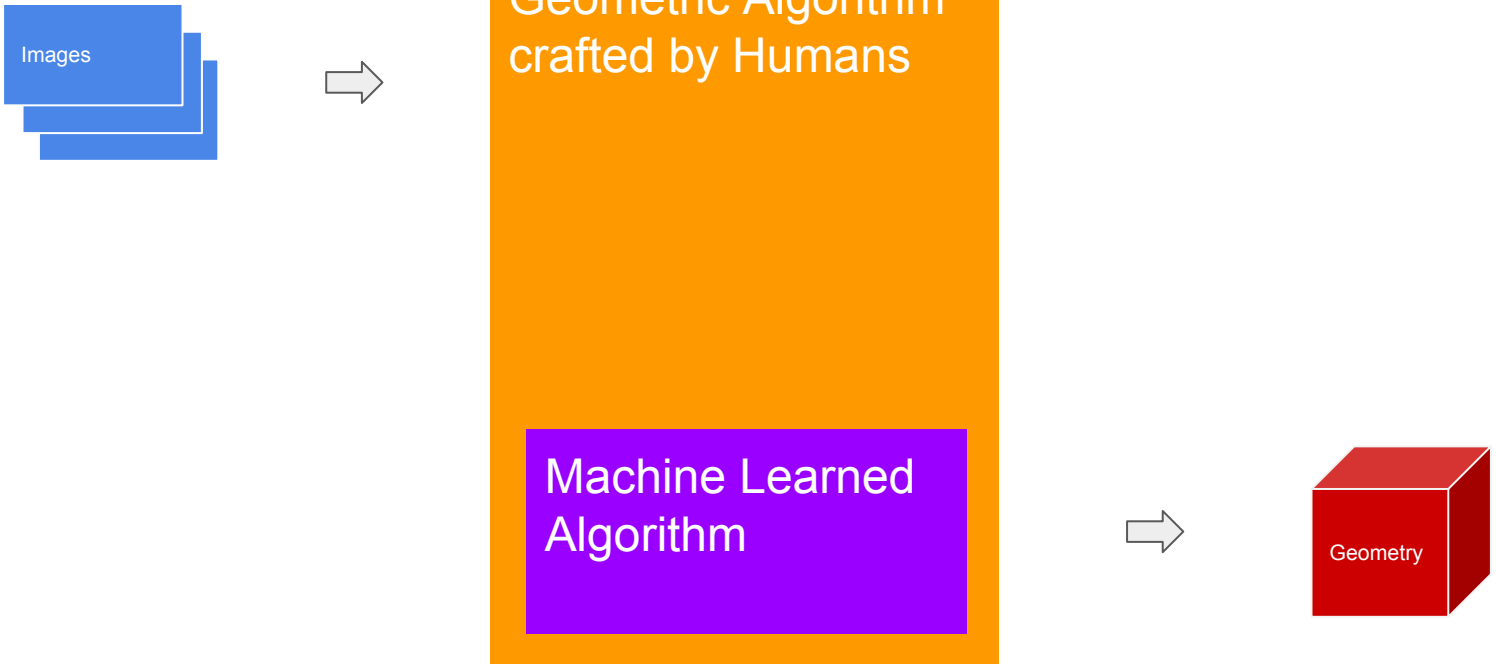


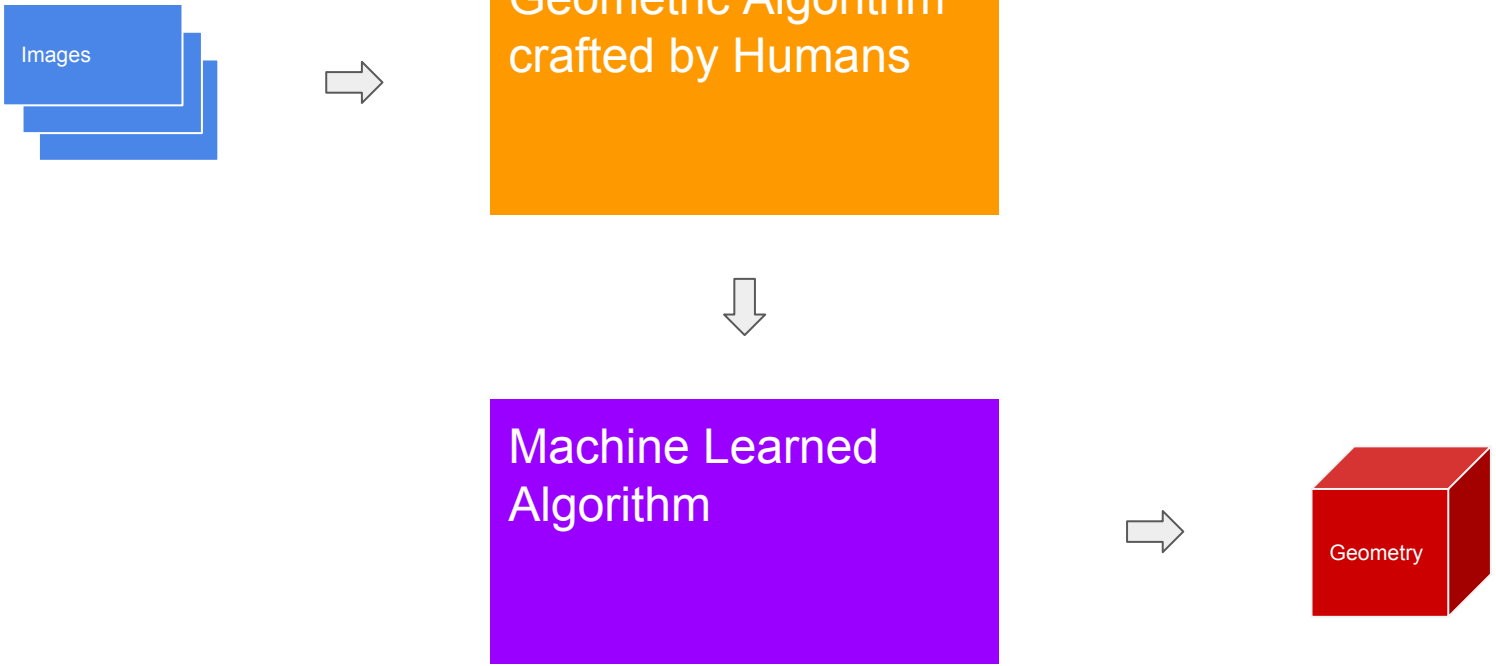
Geometric Algorithm crafted by Humans

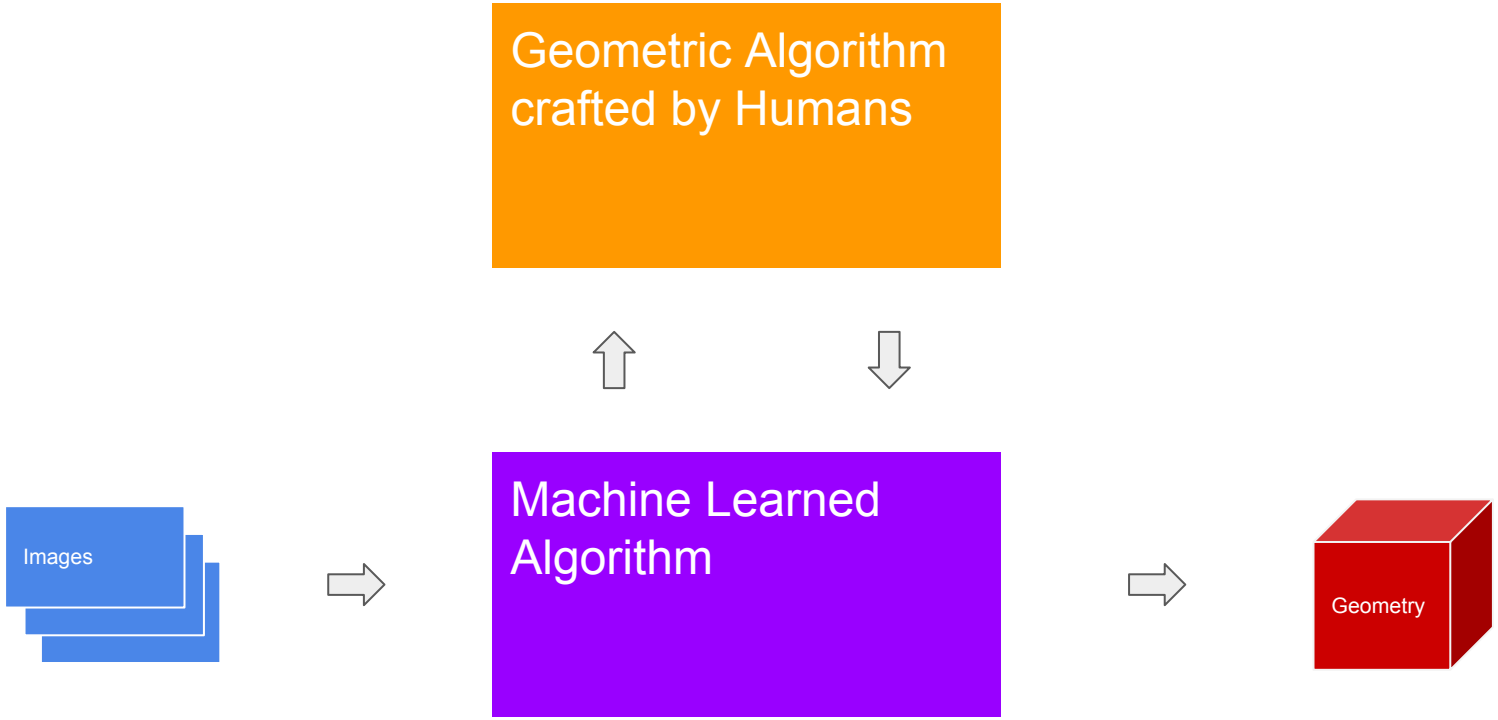



Images
Machine Learned Algorithm
Geometry


Images
Machine Learned Algorithm
Geometric Algorithm crafted by Humans
Geometry


Images
Geometric Algorithm crafted by Humans
Machine Learned Algorithm
Geometry

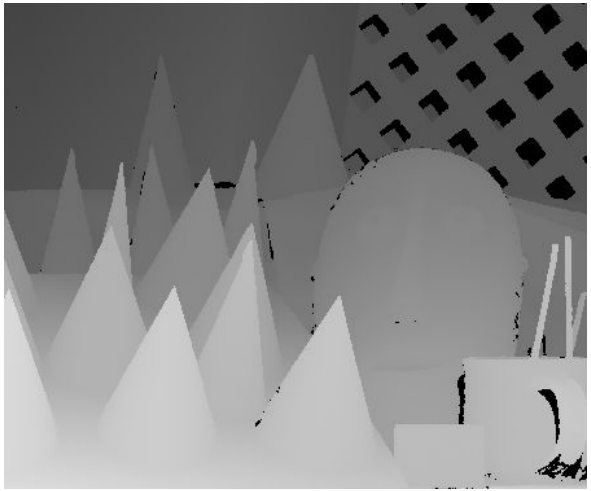

Images
Geometric Algorithm crafted by Humans
Machine Learned Algorithm
Geometry


Geometric Algorithm crafted by Humans
Images
Machine Learned Algorithm
Geometry

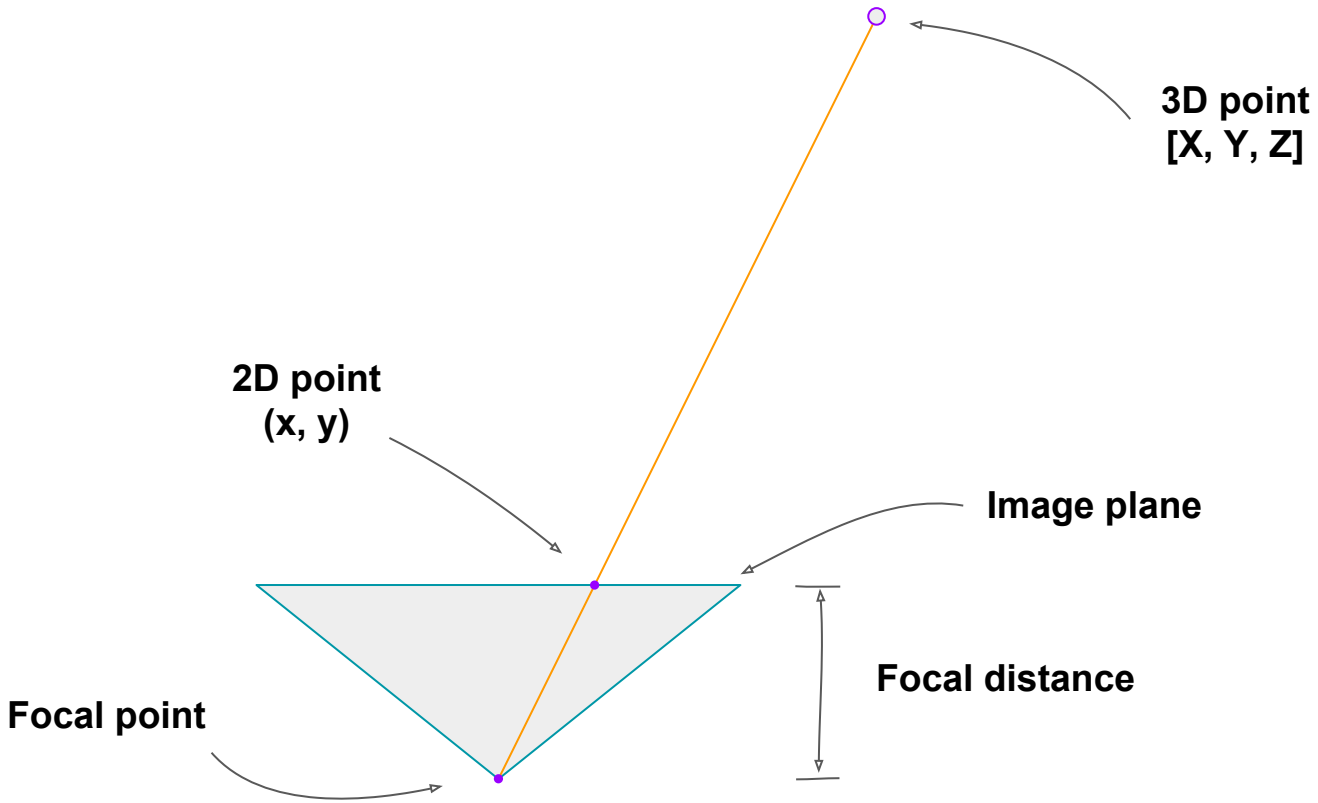
1. Dense stereo correspondence
2. Visual odometry
3. Depth map fusion
4. SLAM

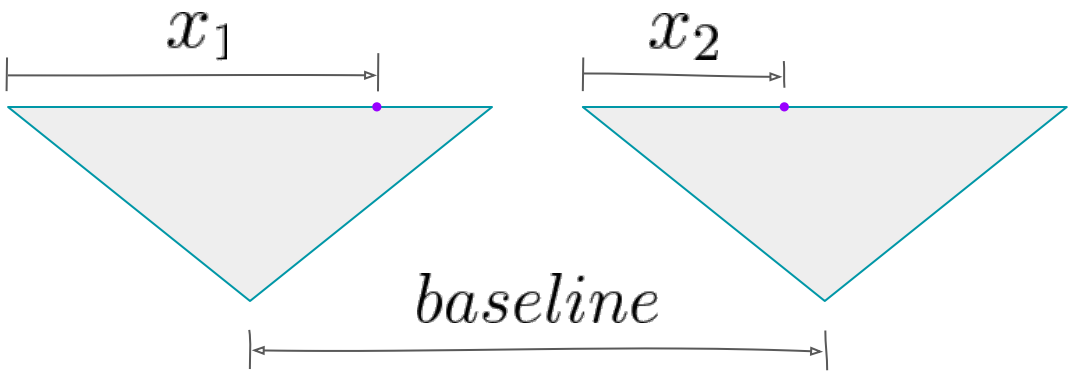

Fisherbrand
Safety Matches

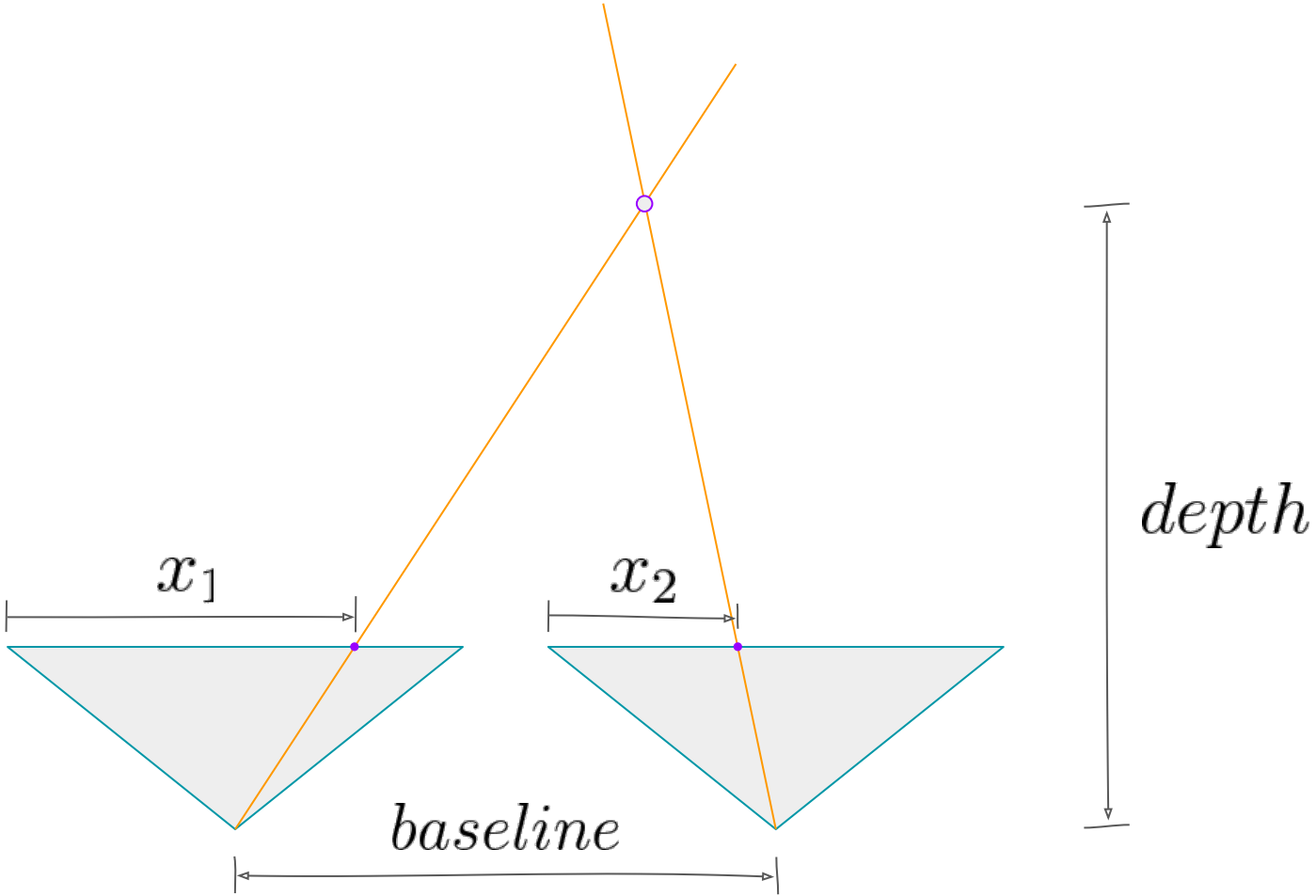

Fisherbrand
Safety Matches

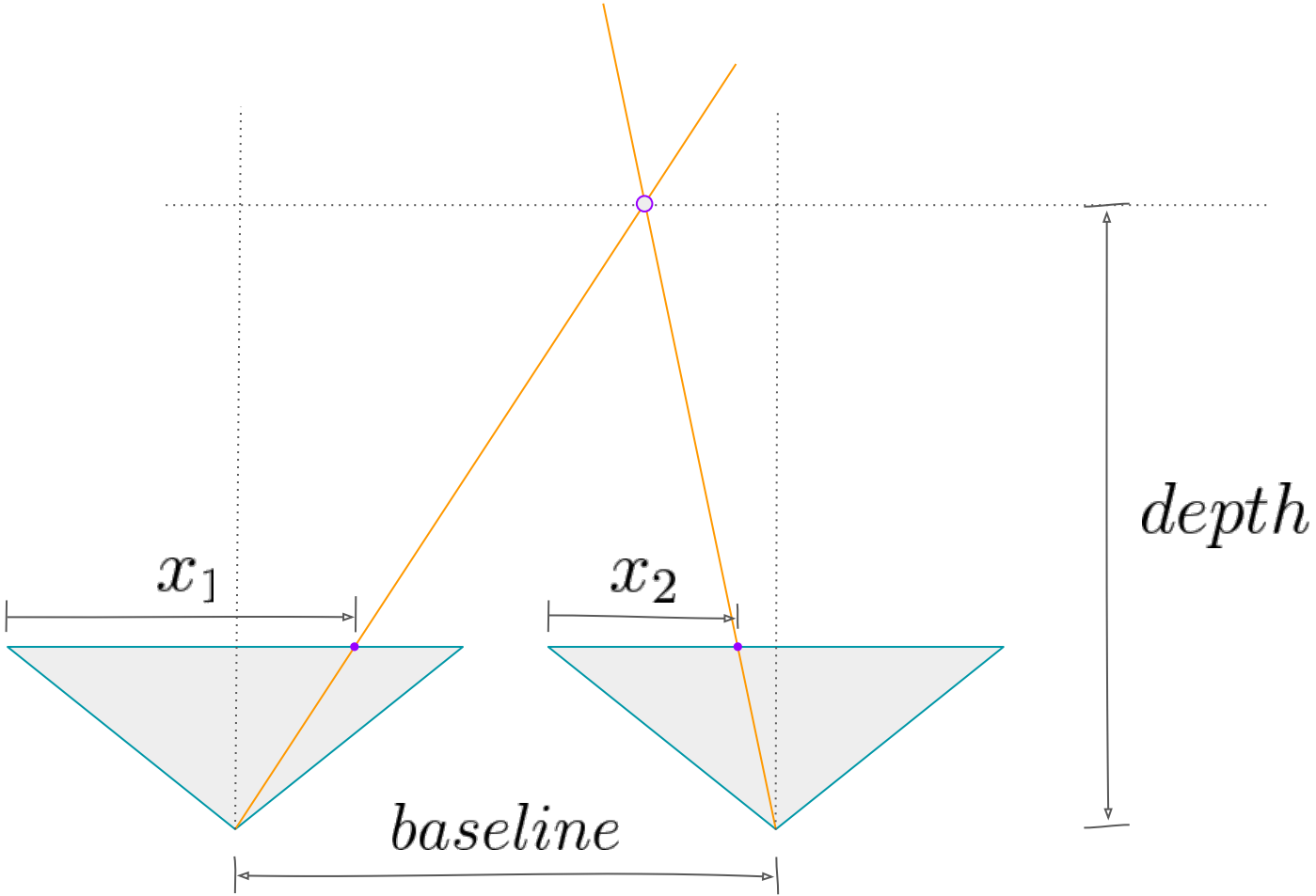


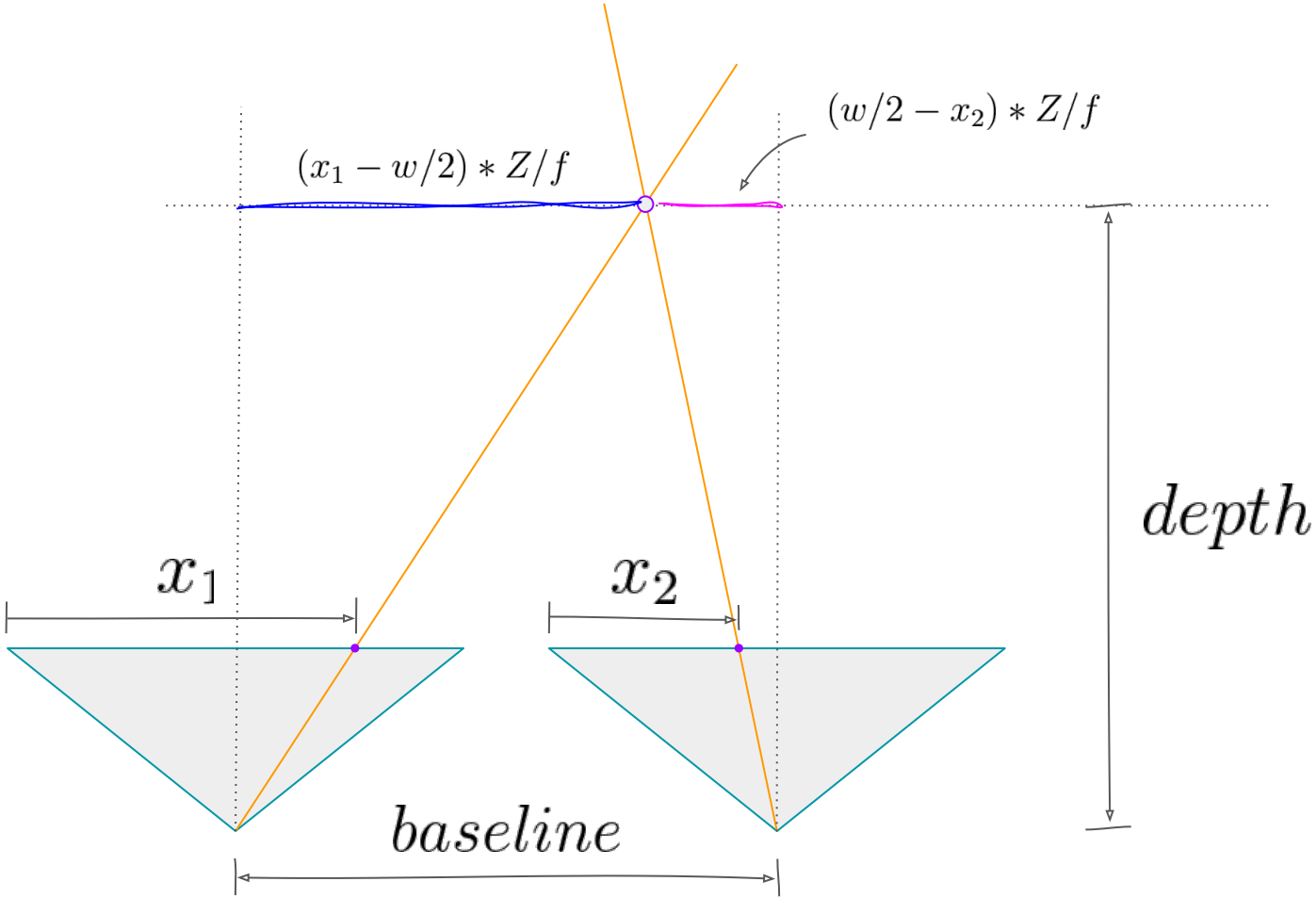
Pinhole camera model



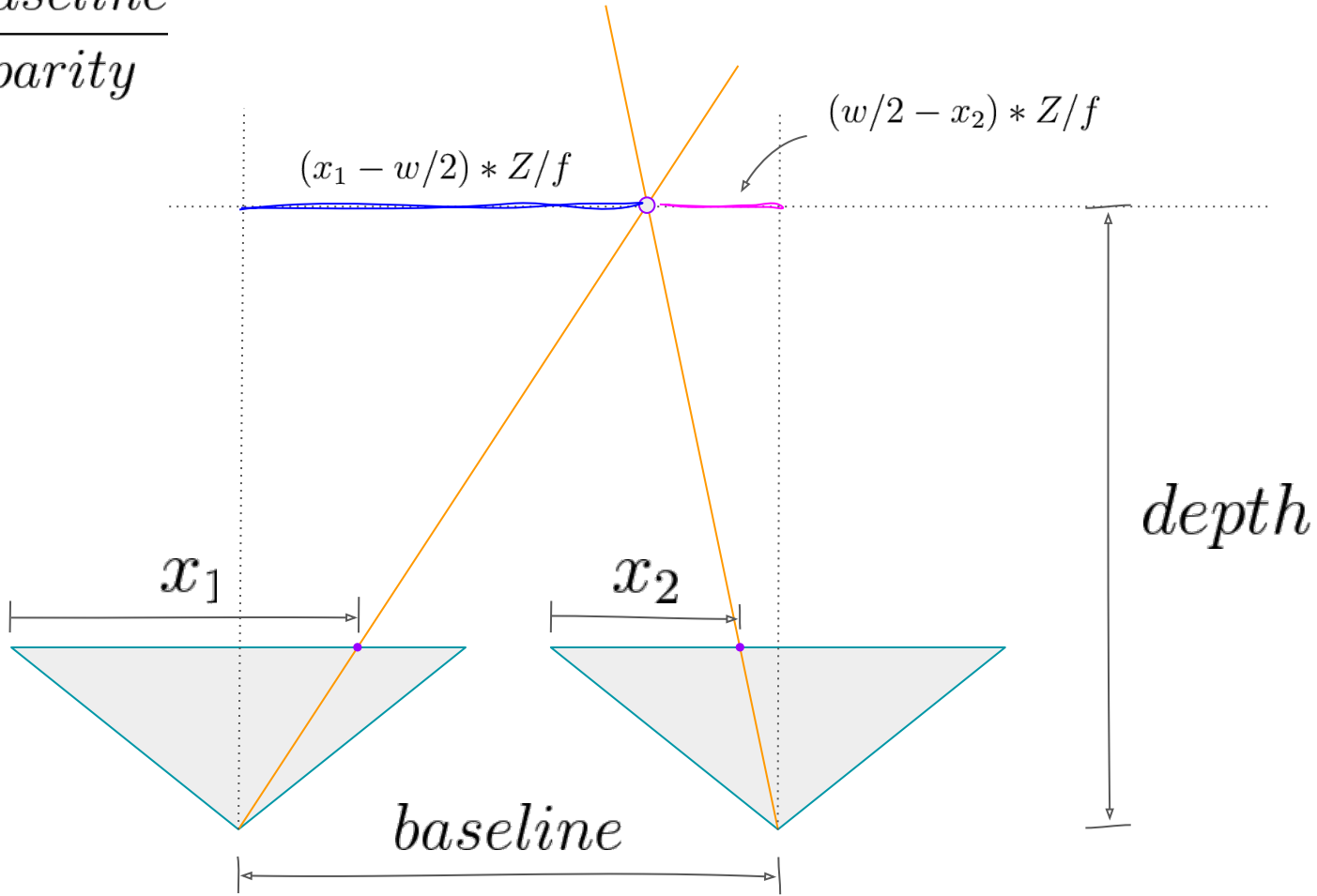

x_1
x_2
$baseline$


x_1
x_2
$depth$
$baseline$


x_1
x_2
$depth$
$baseline$

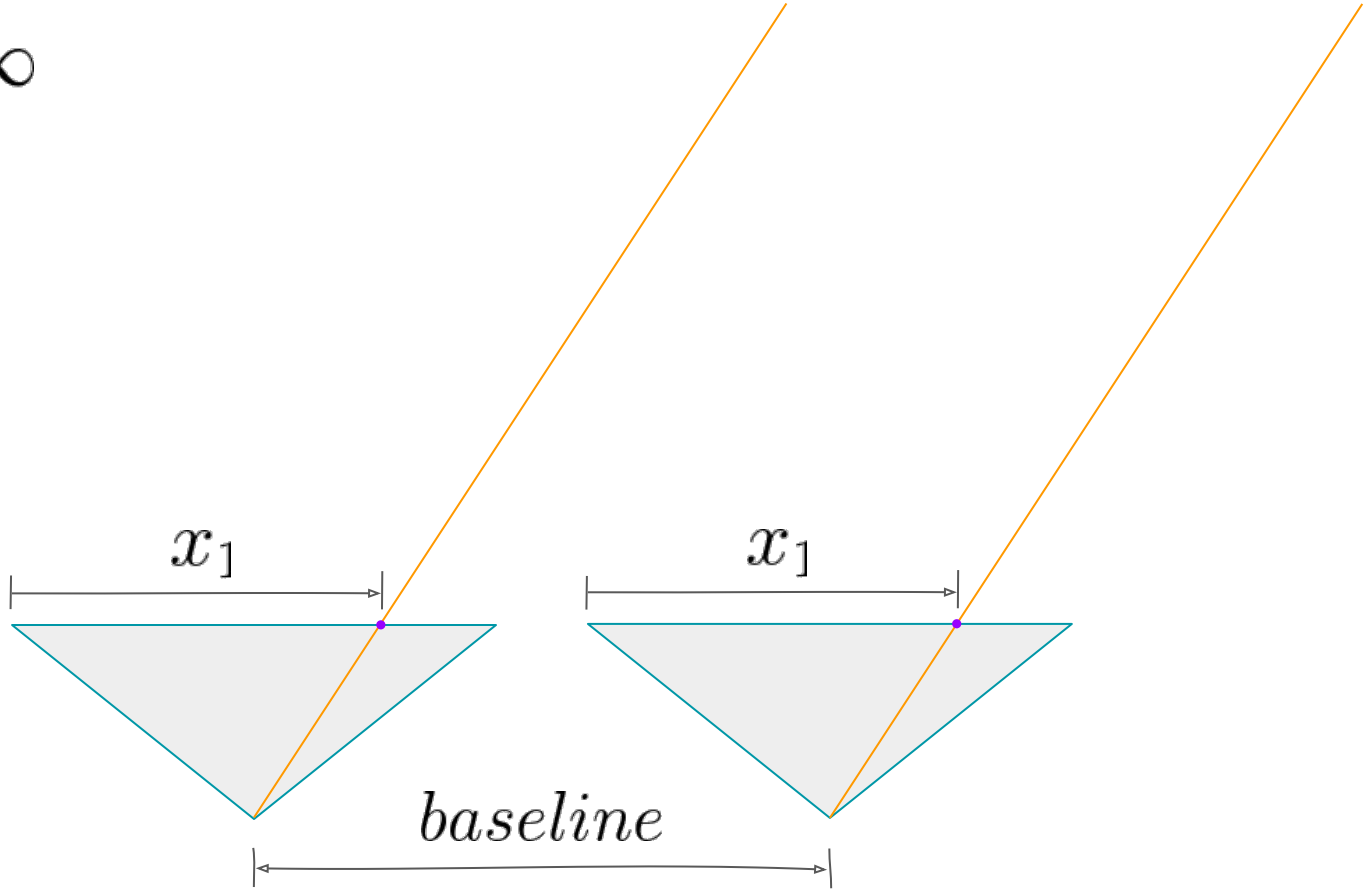

$(x_1 - w/2) * Z/f$
$(w/2 - x_2) * Z/f$
x_1
x_2
$depth$
$baseline$

$$depth = \frac{f * baseline}{disparity}$$



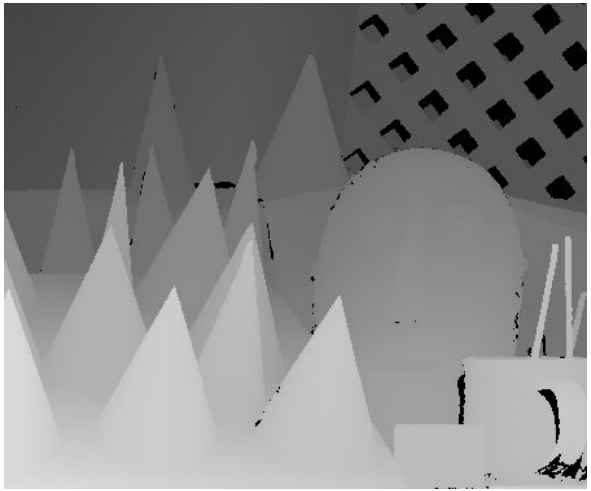
$$disparity = 0$$

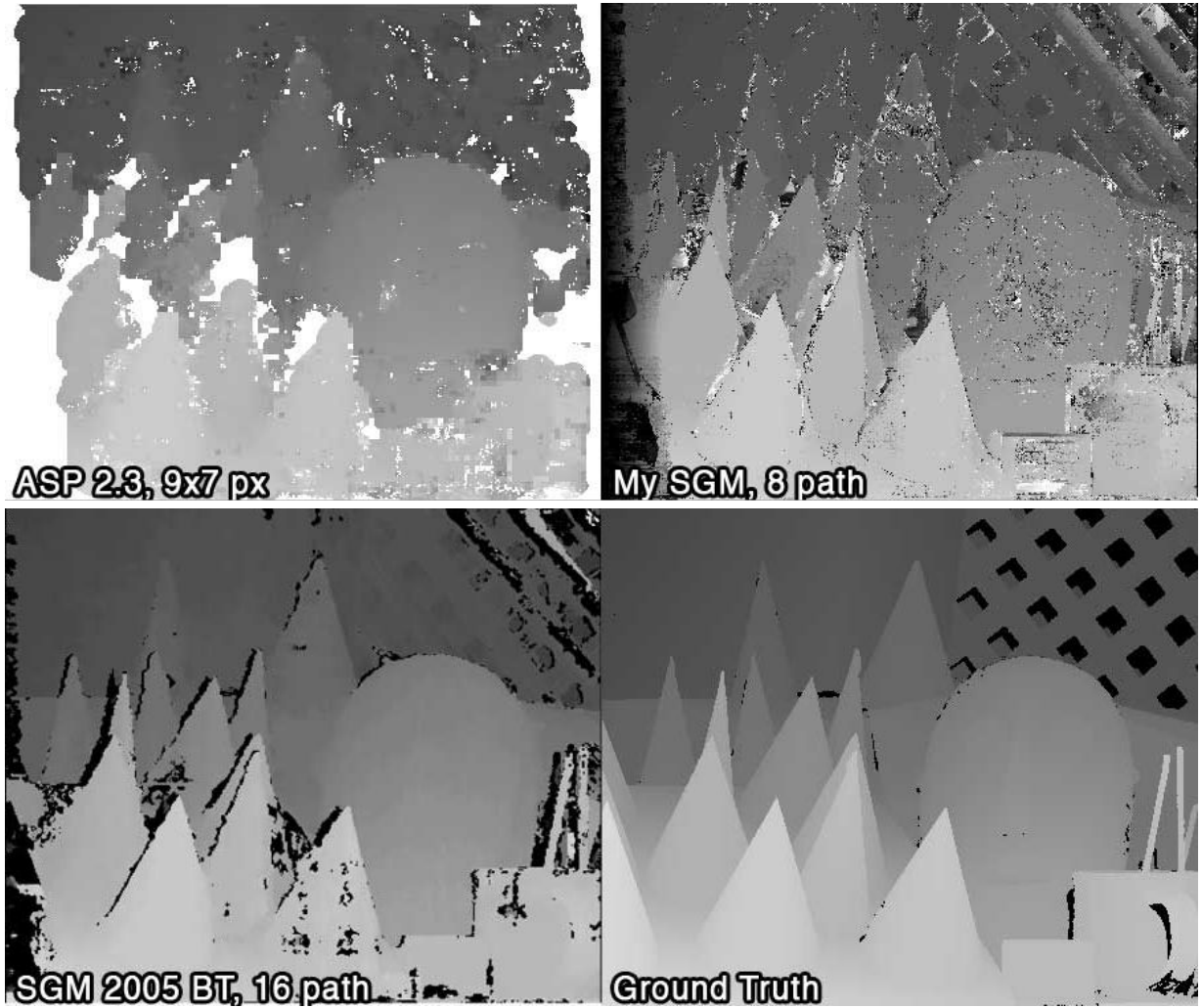
$$depth = \infty$$

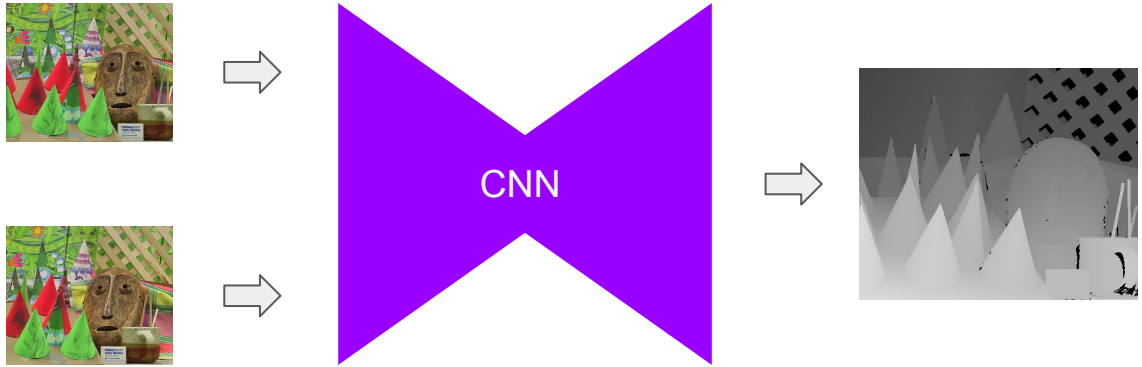


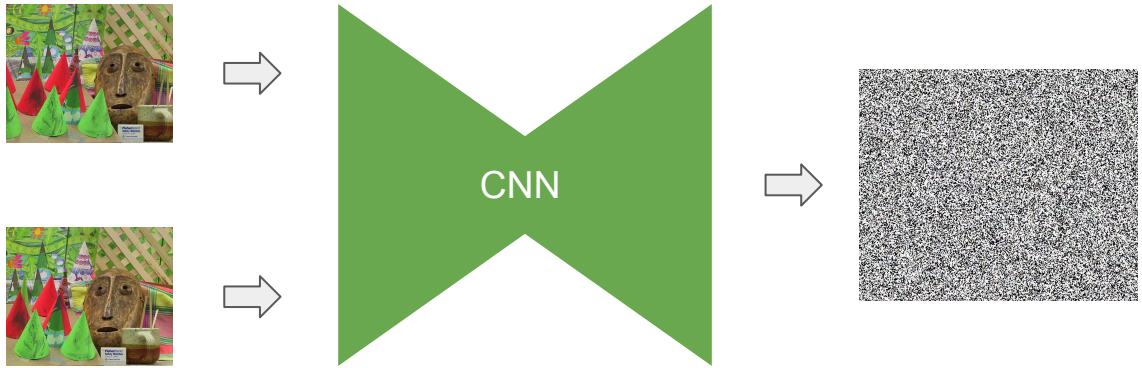

Fisherbrand
Safety Matches

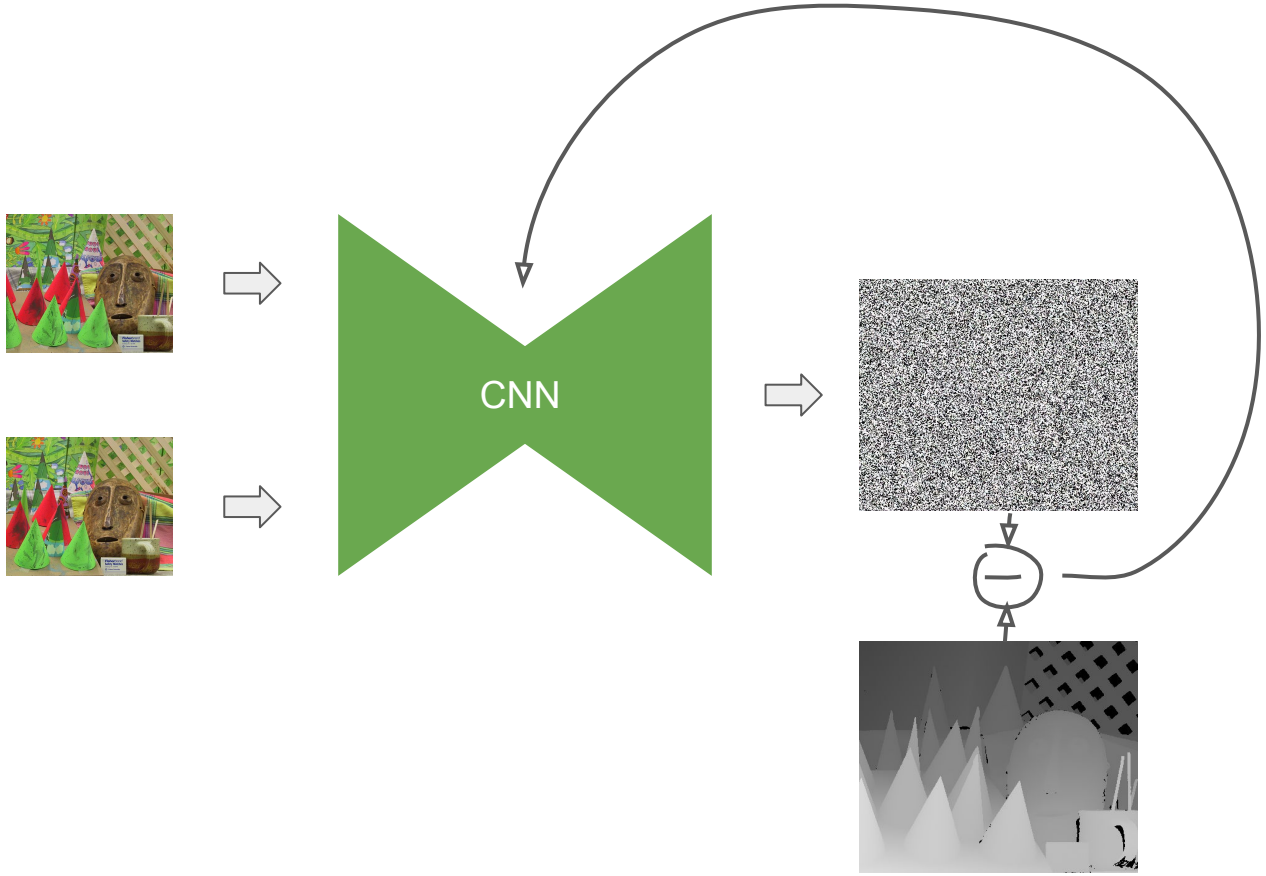

Fisherbrand
Safety Matches

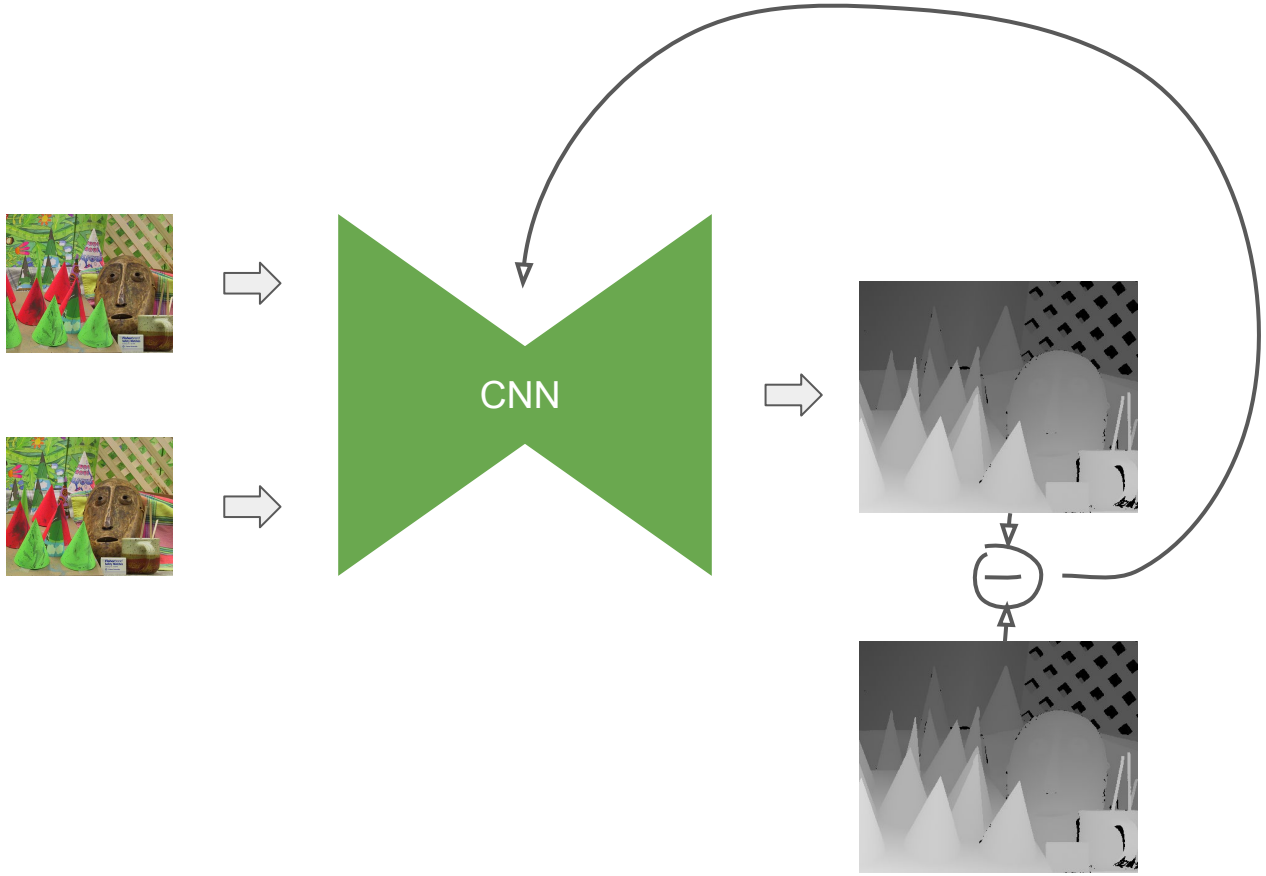


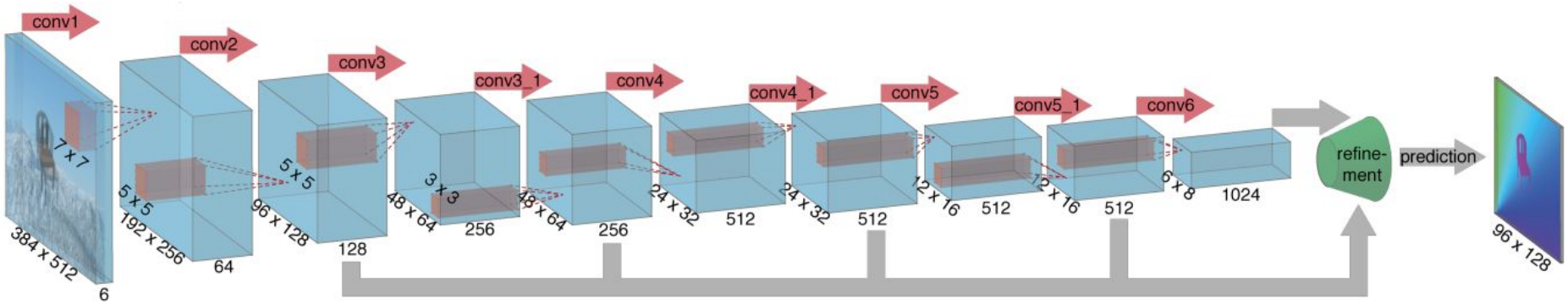

ASP 2.3, 9x7 px
My SGM, 8 path
SGM 2005 BT, 16 path
Ground Truth

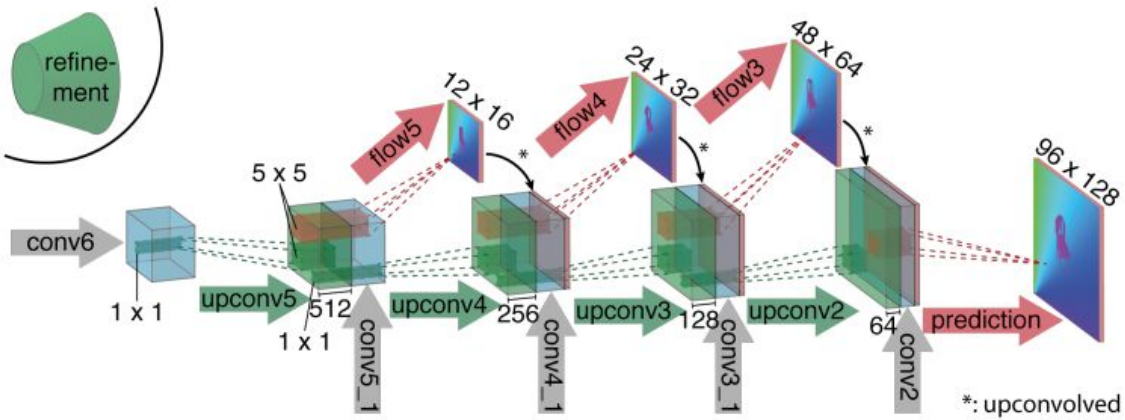

CNN

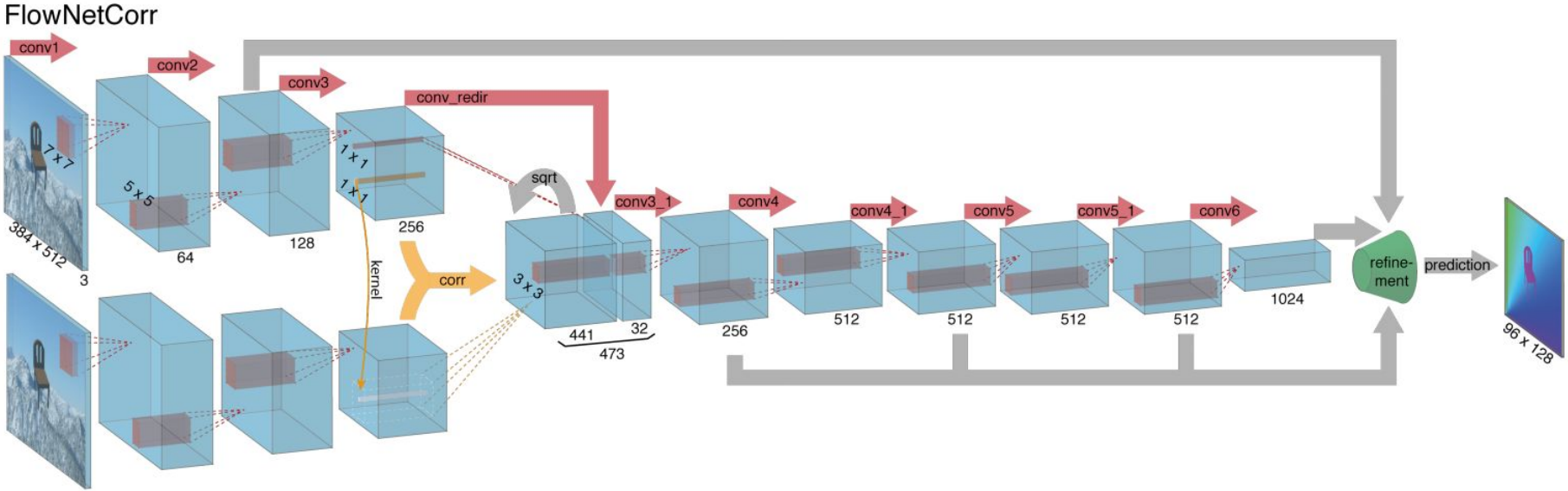

CNN

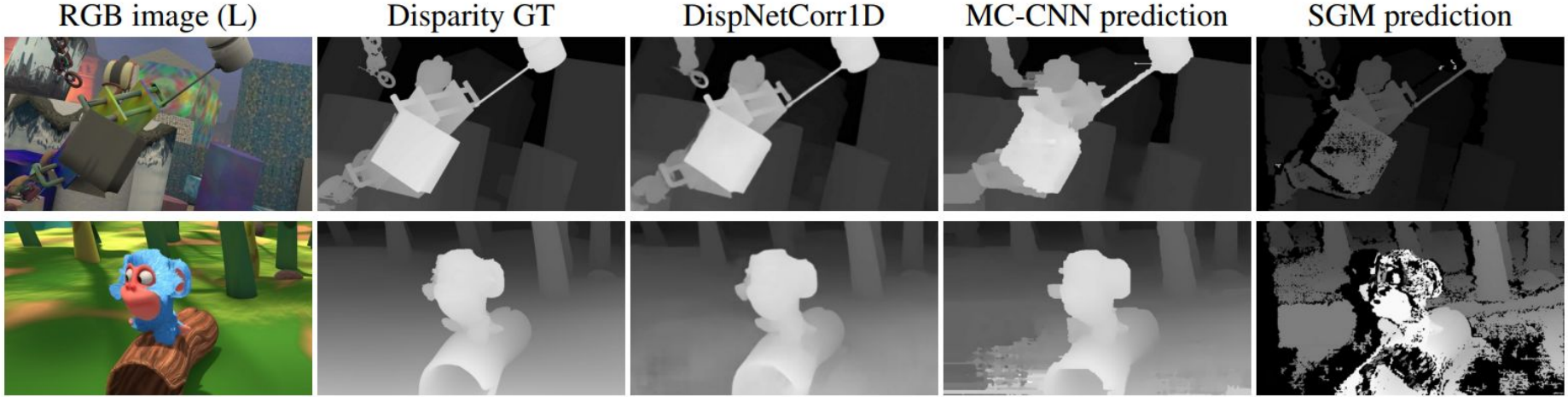

CNN


CNN


conv1
conv2
conv3
conv3_1
conv4
conv4_1
conv5
conv5_1
conv6
7 x 7
5 x 5
5 x 5
3 x 3
384 x 512
6
192 x 256
64
96 x 128
128
48 x 64
256
48 x 64
256
24 x 32
512
24 x 32
512
12 x 16
512
12 x 16
512
6 x 8
1024
refine-
ment
prediction
96 x 128


refine-
ment
conv6
1 x 1
upconv5
5 x 5
512
1 x 1
conv5_1
flow5
12 x 16
*
upconv4
256
conv4_1
flow4
24 x 32
*
upconv3
128
conv3_1
flow3
48 x 64
*
upconv2
64
conv2
prediction
96 x 128
*: upconvolved


FlowNetCorr
conv1
conv2
conv3
conv_redir
7 x 7
5 x 5
1 x 1
1 x 1
384 x 512
3
64
128
256
sqrt
kernel
corr
3 x 3
441
32
473
conv3_1
conv4
conv4_1
conv5
conv5_1
conv6
256
512
512
512
512
1024
refine-
ment
prediction
96 x 128


RGB image (L)
Disparity GT
DispNetCorr1D
MC-CNN prediction
SGM prediction

A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation

Nikolaus Mayer*[1], Eddy Ilg*[1], Philip Häusser*[2], Philipp Fischer*[1†]
[1]University of Freiburg [2]Technical University of Munich
[1]{mayern,ilg,fischer}@cs.uni-freiburg.de [2]haeusser@cs.tum.edu

Daniel Cremers
Technical University of Munich
cremers@tum.de

Alexey Dosovitskiy, Thomas Brox
University of Freiburg
{dosovits,brox}@cs.uni-freiburg.de

Abstract

Recent work has shown that optical flow estimation can be formulated as a supervised learning task and can be successfully solved with convolutional networks. Training of the so-called FlowNet was enabled by a large synthetically generated dataset. The present paper extends the concept of optical flow estimation via convolutional networks to disparity and scene flow estimation. To this end, we propose three synthetic stereo video datasets with sufficient realism, variation, and size to successfully train large networks. Our datasets are the first large-scale datasets to enable training and evaluating scene flow methods. Besides the datasets, we present a convolutional network for real-time disparity estimation that provides state-of-the-art results. By combining a flow and disparity estimation network and training it jointly, we demonstrate the first scene flow estimation with a convolutional network.

Figure 1. Our datasets provide over 35 000 stereo frames with dense ground truth for *optical flow*, *disparity* and *disparity change*, as well as other data such as object segmentation.

A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation

Nikolaus Mayer*[1], Eddy Ilg*[1], Philip Häusser*[2], Philipp Fischer*[1†]
[1]University of Freiburg [2]Technical University of Munich
[1]{mayern,ilg,fischer}@cs.uni-freiburg.de [2]haeusser@cs.tum.edu

Daniel Cremers
Technical University of Munich
cremers@tum.de

Alexey Dosovitskiy, Thomas Brox
University of Freiburg
{dosovits,brox}@cs.uni-freiburg.de

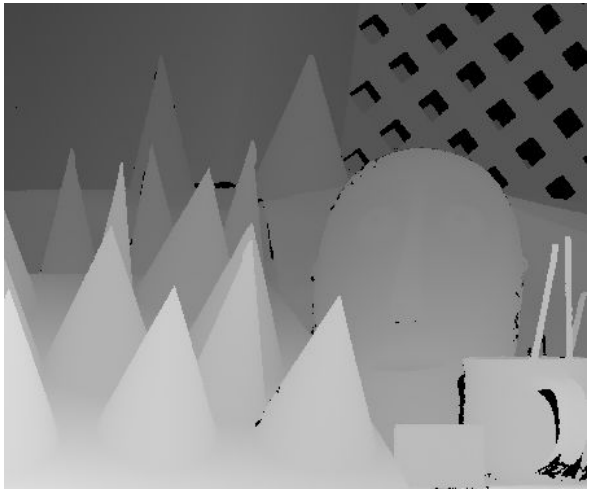
Abstract

Recent work has shown that optical flow estimation can be formulated as a supervised learning task and can be successfully solved with convolutional networks. Training of the so-called FlowNet was enabled by a large synthetically generated dataset. The present paper extends the concept of optical flow estimation via convolutional networks to disparity and scene flow estimation. To this end, we propose three synthetic stereo video datasets with sufficient realism, variation, and size to successfully train large networks. Our datasets are the first large-scale datasets to enable training and evaluating scene flow methods. Besides the datasets, we present a convolutional network for real-time disparity estimation that provides state-of-the-art results. By combining a flow and disparity estimation network and training it jointly, we demonstrate the first scene flow estimation with a convolutional network.

Figure 1. Our datasets provide over 35 000 stereo frames with dense ground truth for *optical flow*, *disparity* and *disparity change*, as well as other data such as object segmentation.

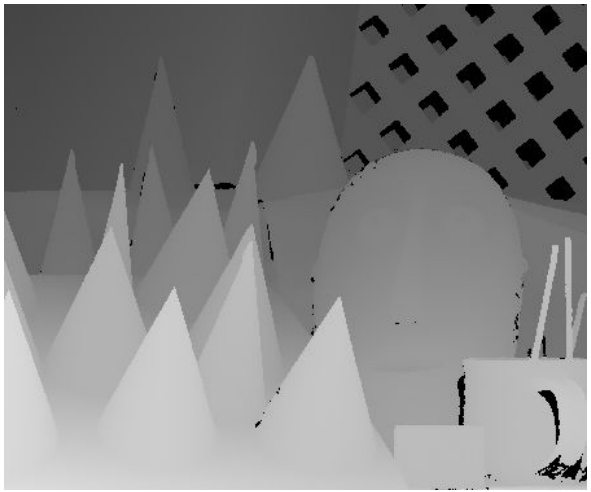

Fisherbrand
Safety Matches


Fisherbrand
Safety Matches



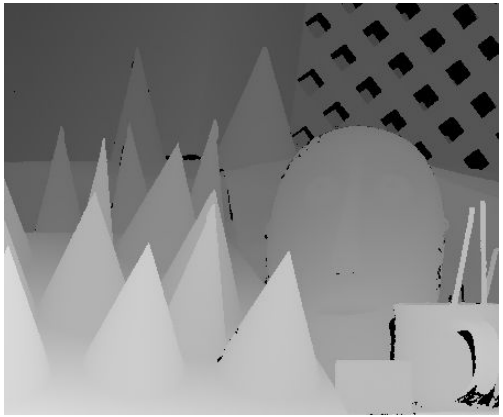

Fisherbrand
Safety Matches

?



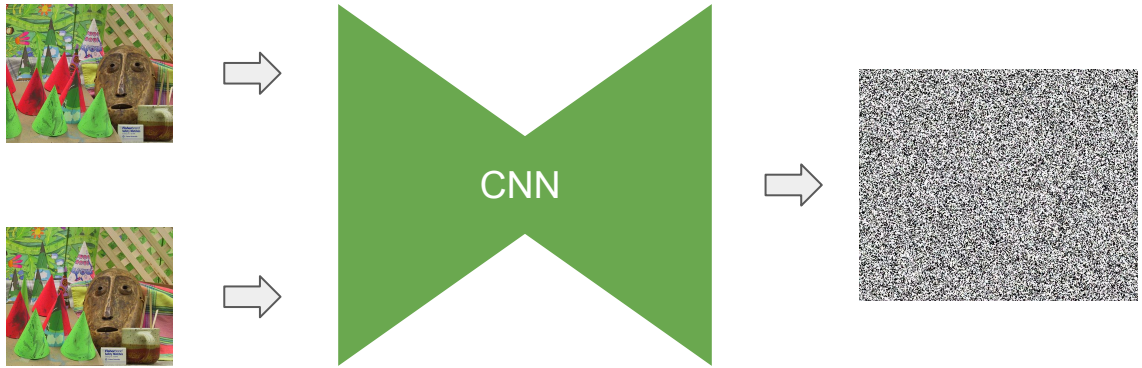

Fisherbrand
Safety Matches

+

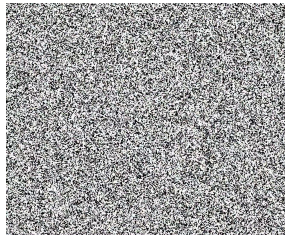


=

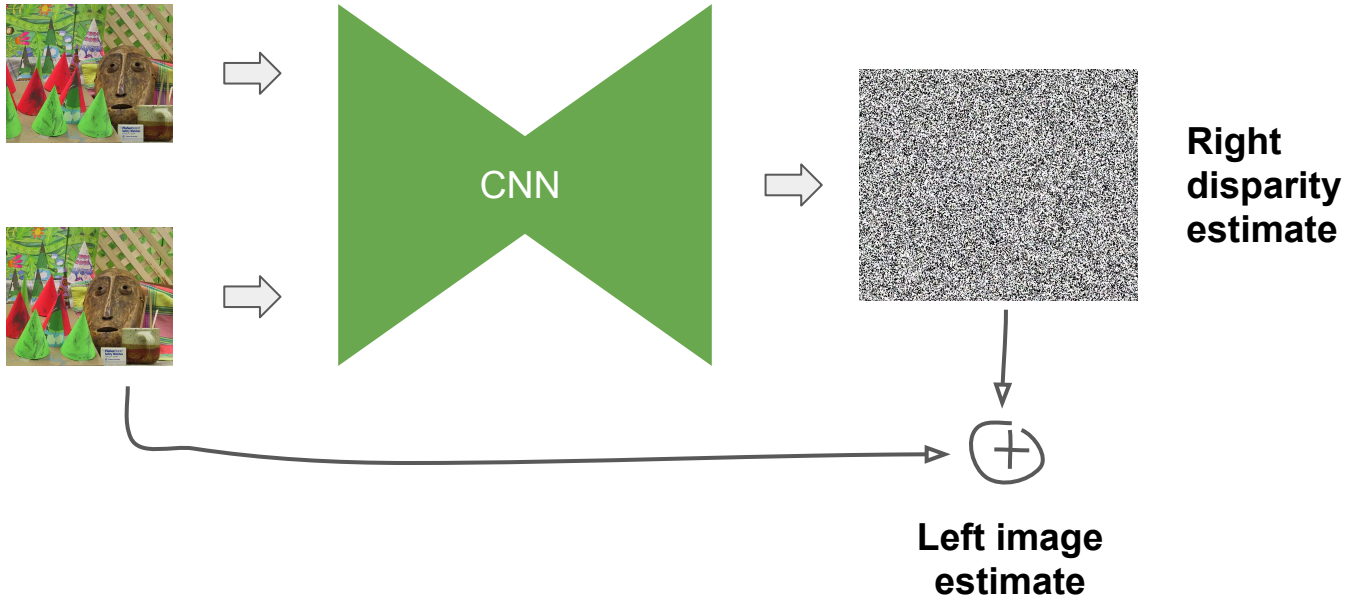

Fisherbrand
Safety Matches

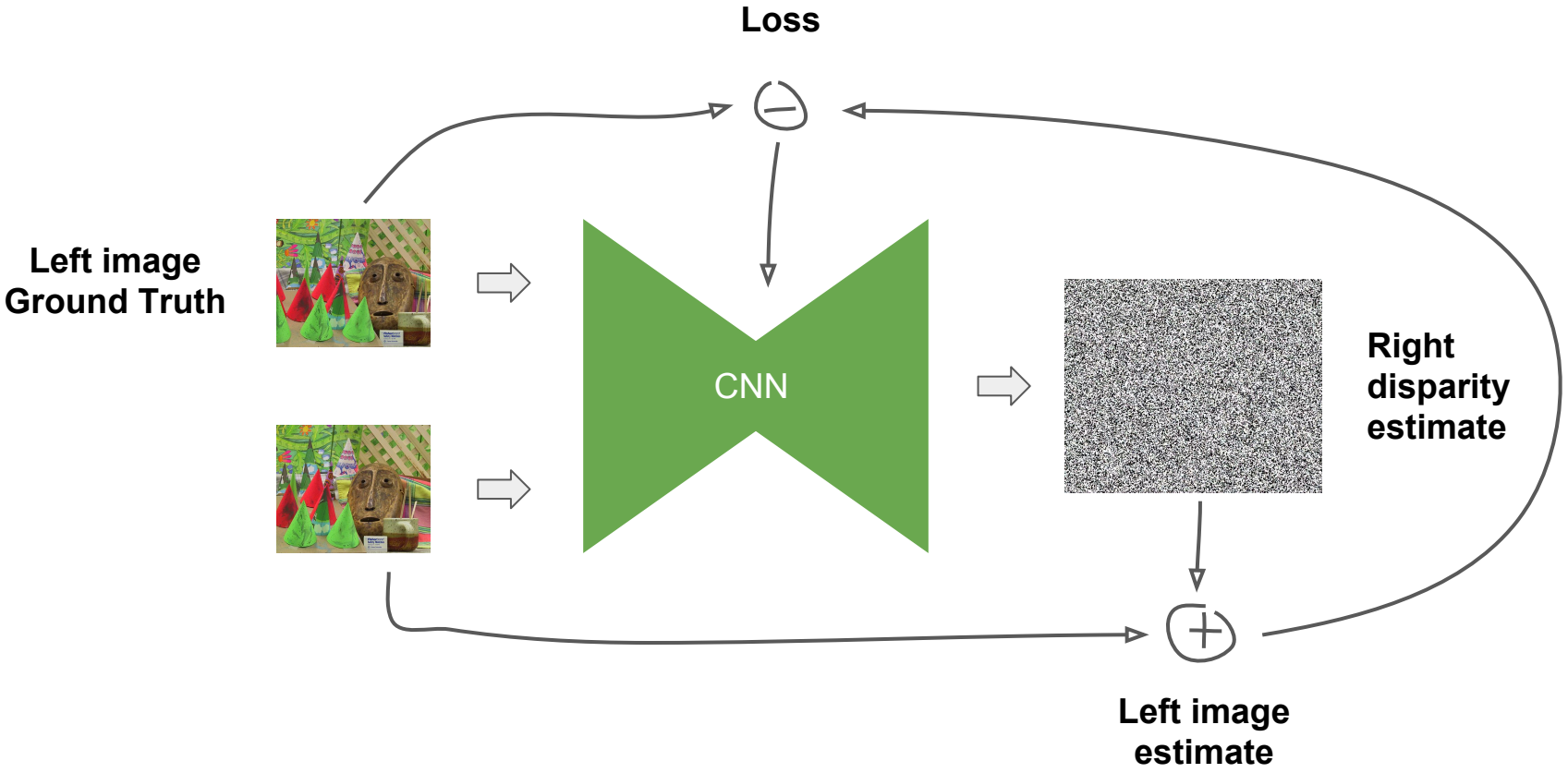

CNN

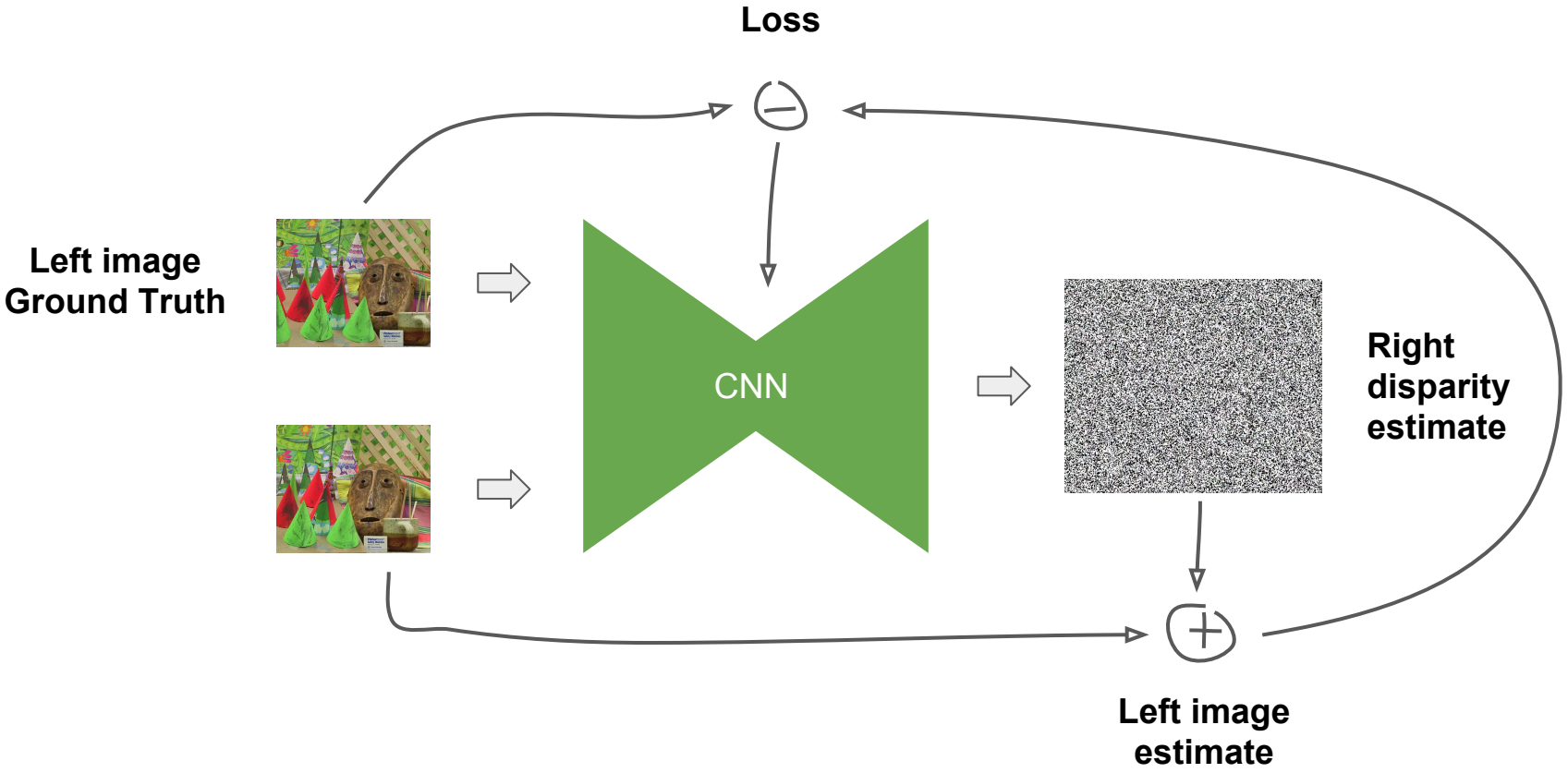
CNN

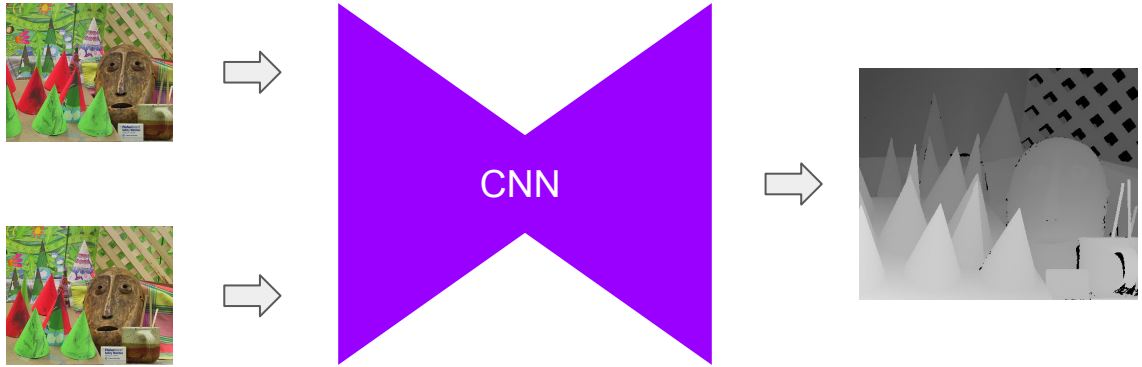


Right disparity estimate


CNN
Right disparity estimate
Left image estimate


Loss
Left image
Ground Truth
CNN
Right
disparity
estimate
Left image
estimate


Loss
Left image
Ground Truth
CNN
Right
disparity
estimate
Left image
estimate


CNN

Unsupervised [redacted] Depth Estimation with Left-Right Consistency

Clément Godard Oisin Mac Aodha Gabriel J. Brostow
University College London
http://visual.cs.ucl.ac.uk/pubs/[redacted]/

Abstract

Learning based methods have shown very promising results for the task of [redacted] However, most existing approaches treat depth prediction as a supervised regression problem and as a result, require vast quantities of corresponding ground truth depth data for training. Just recording quality depth data in a range of environments is a challenging problem. In this paper, we innovate beyond existing approaches, replacing the use of explicit depth data during training with easier-to-obtain binocular stereo footage.

We propose a novel training objective that enables our convolutional neural network to learn to perform single image depth estimation, despite the absence of ground truth depth data. Exploiting epipolar geometry constraints, we generate disparity images by training our network with an image reconstruction loss. We show that solving for image reconstruction alone results in poor quality depth images. To overcome this problem, we propose a novel training loss that enforces consistency between the disparities produced relative to both the left and right

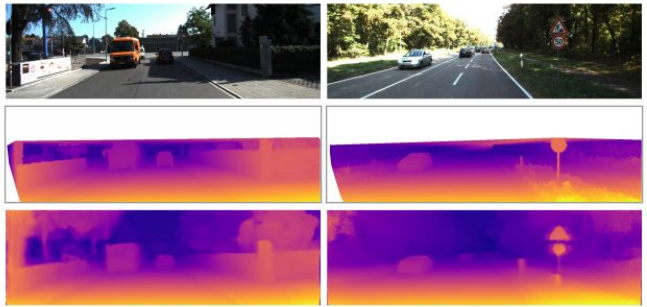
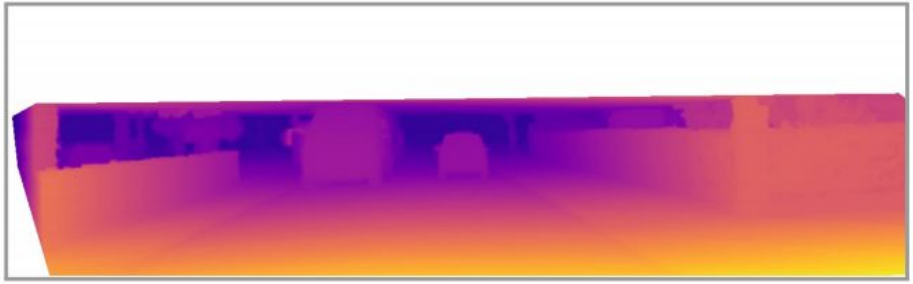
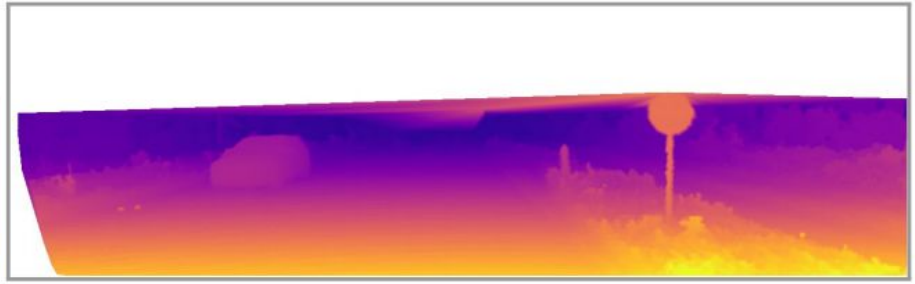
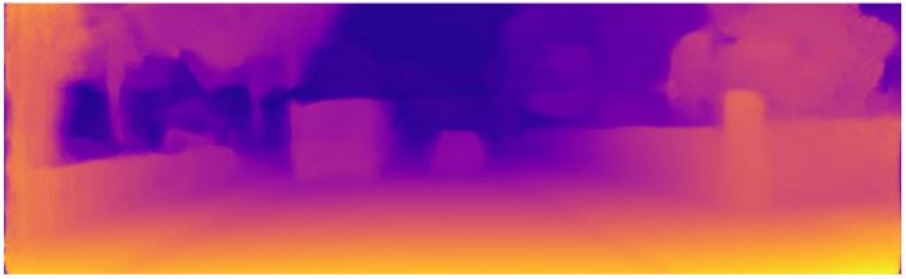
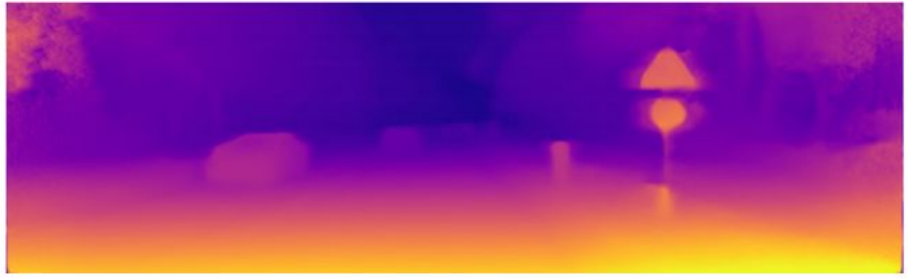


Figure 1. Our depth prediction results on KITTI 2015. Top to bottom: input image, ground truth disparities, and our result. Our method is able to estimate depth for thin structures such as street signs and poles.

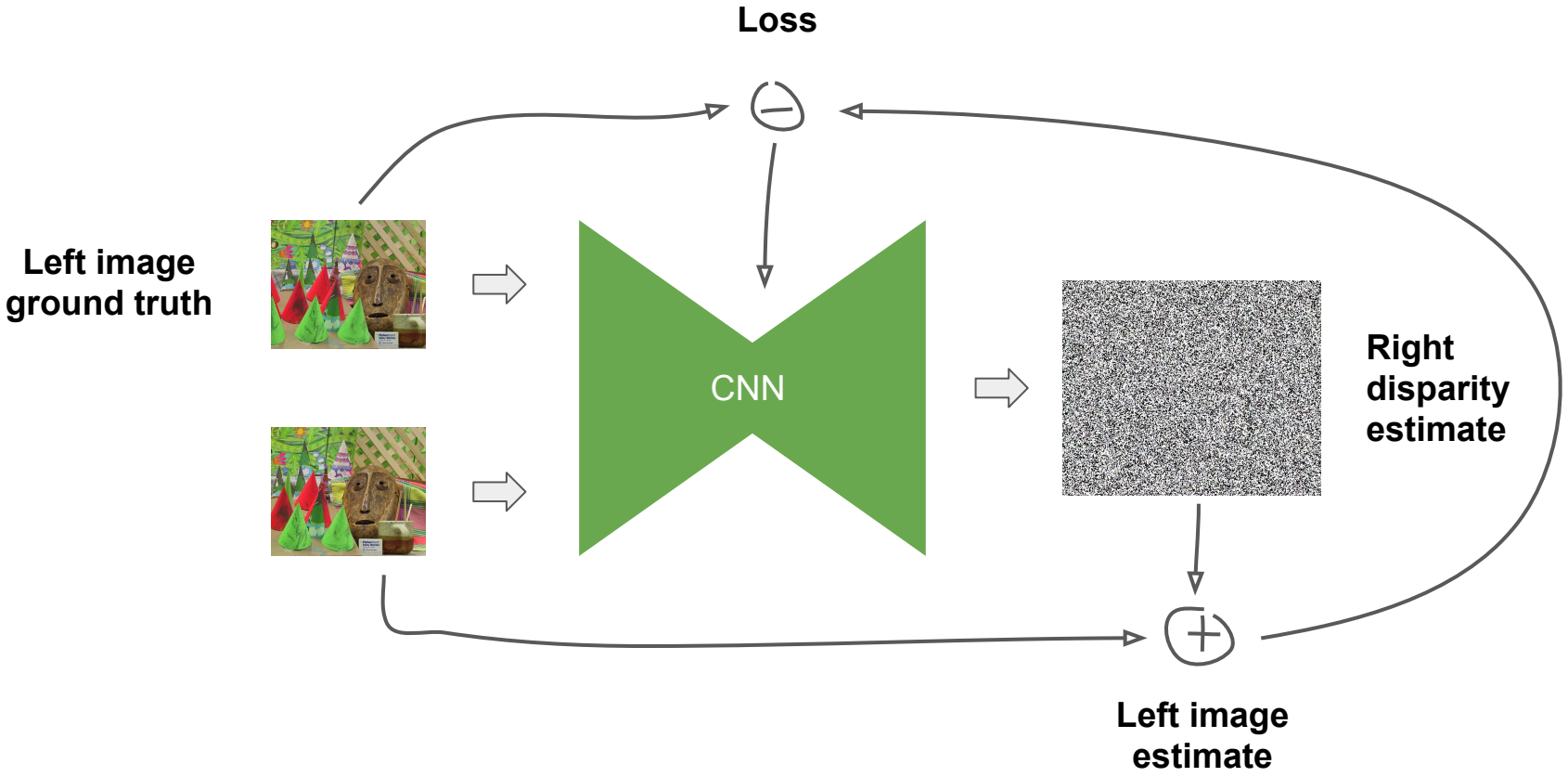
have been restricted to scenes where large image collections and their corresponding pixel depths are available.

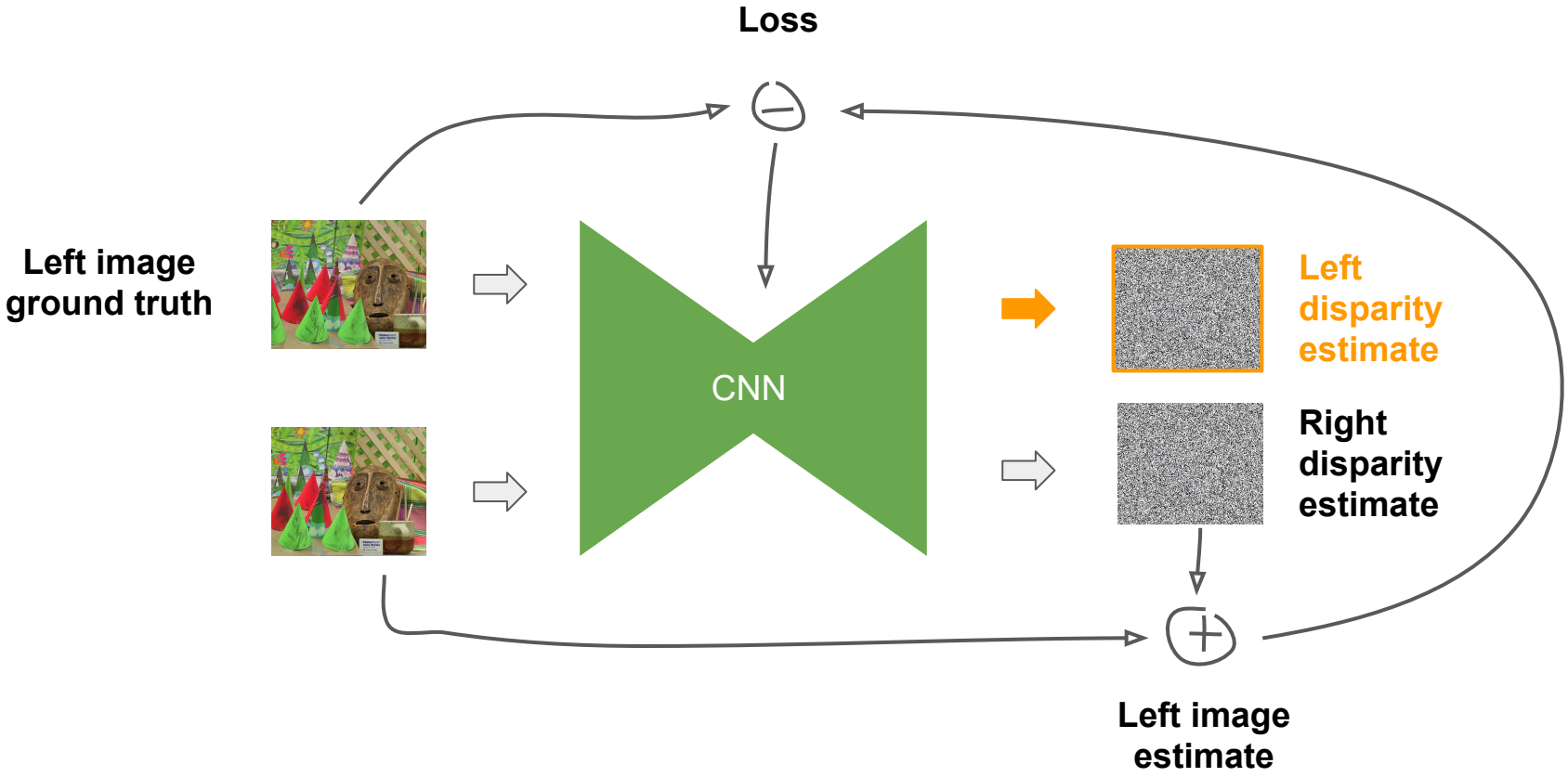
Understanding the shape of a scene from a single image, independent of its appearance, is a fundamental problem in machine perception. There are many applications such as synthetic object insertion in computer graphics [29], synthetic

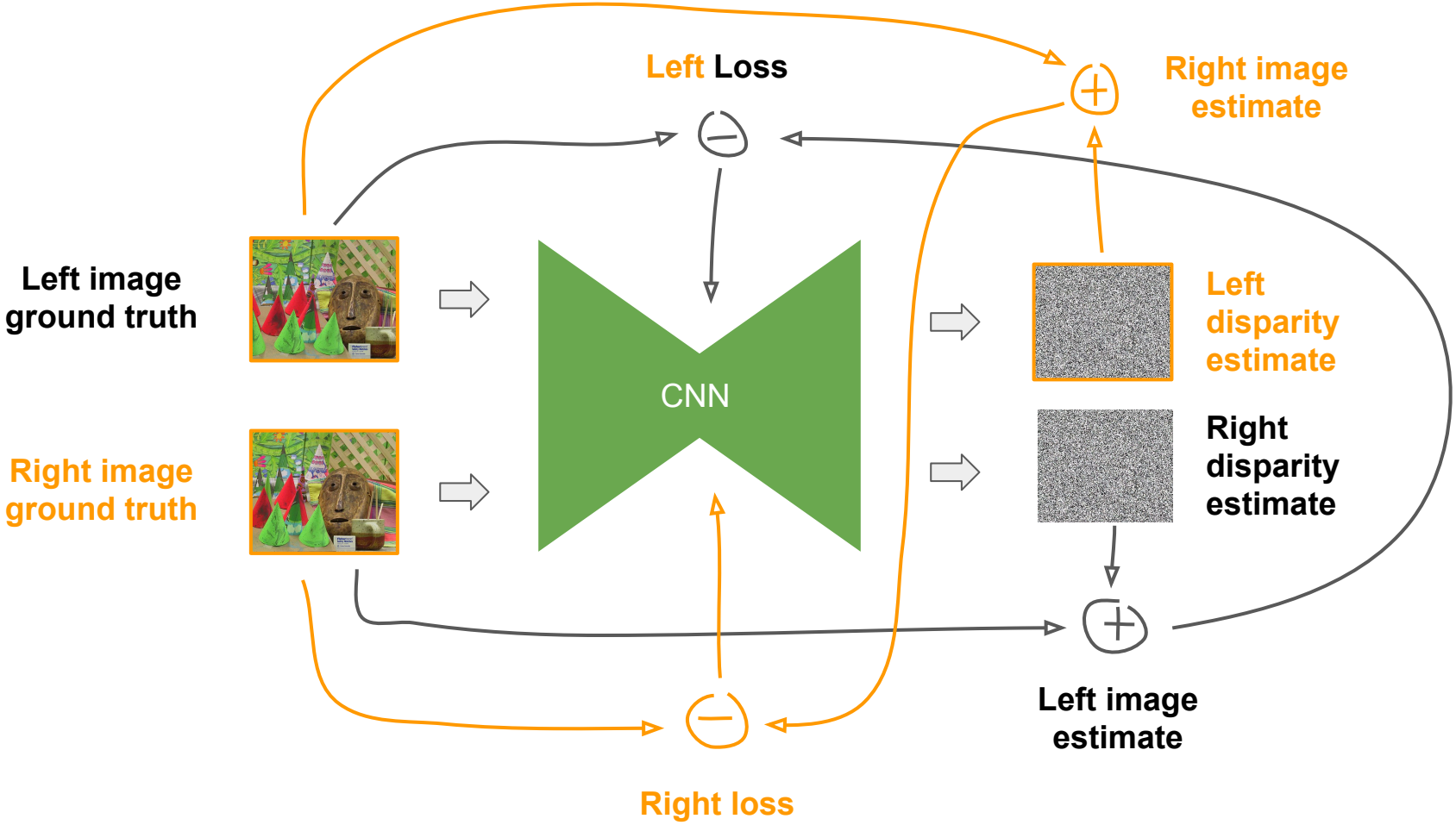


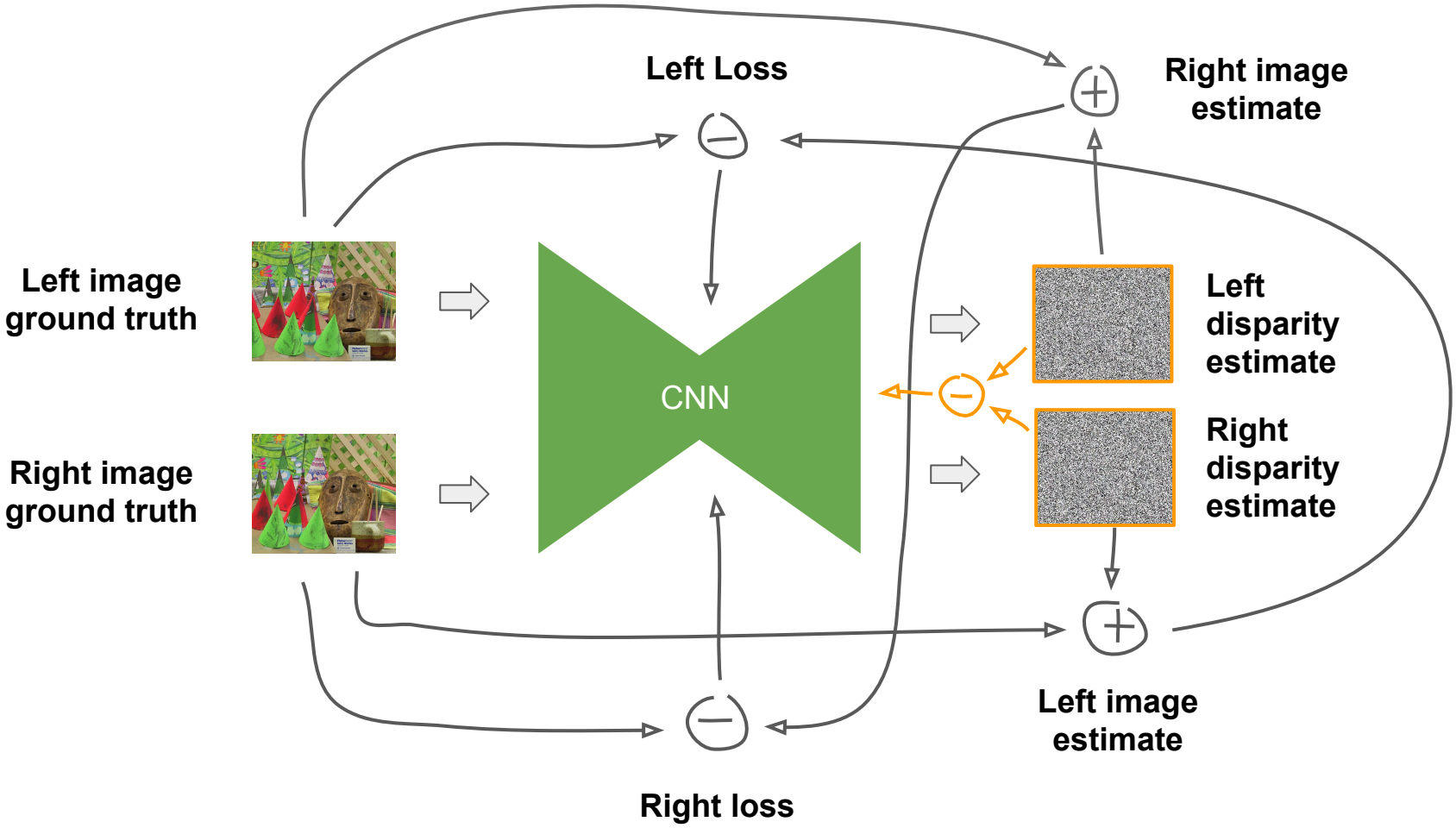





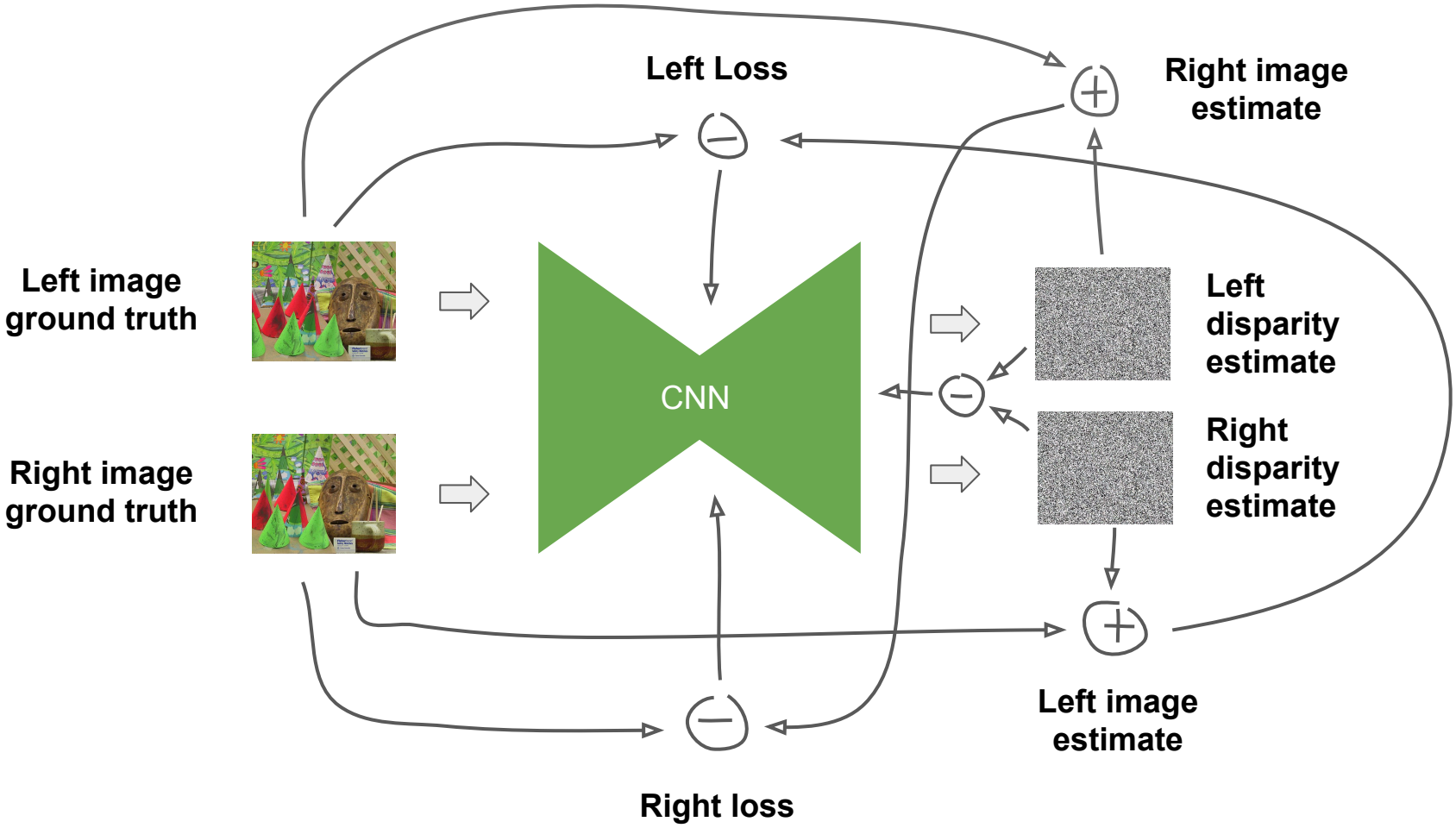

TO LET
0208 7405 4545

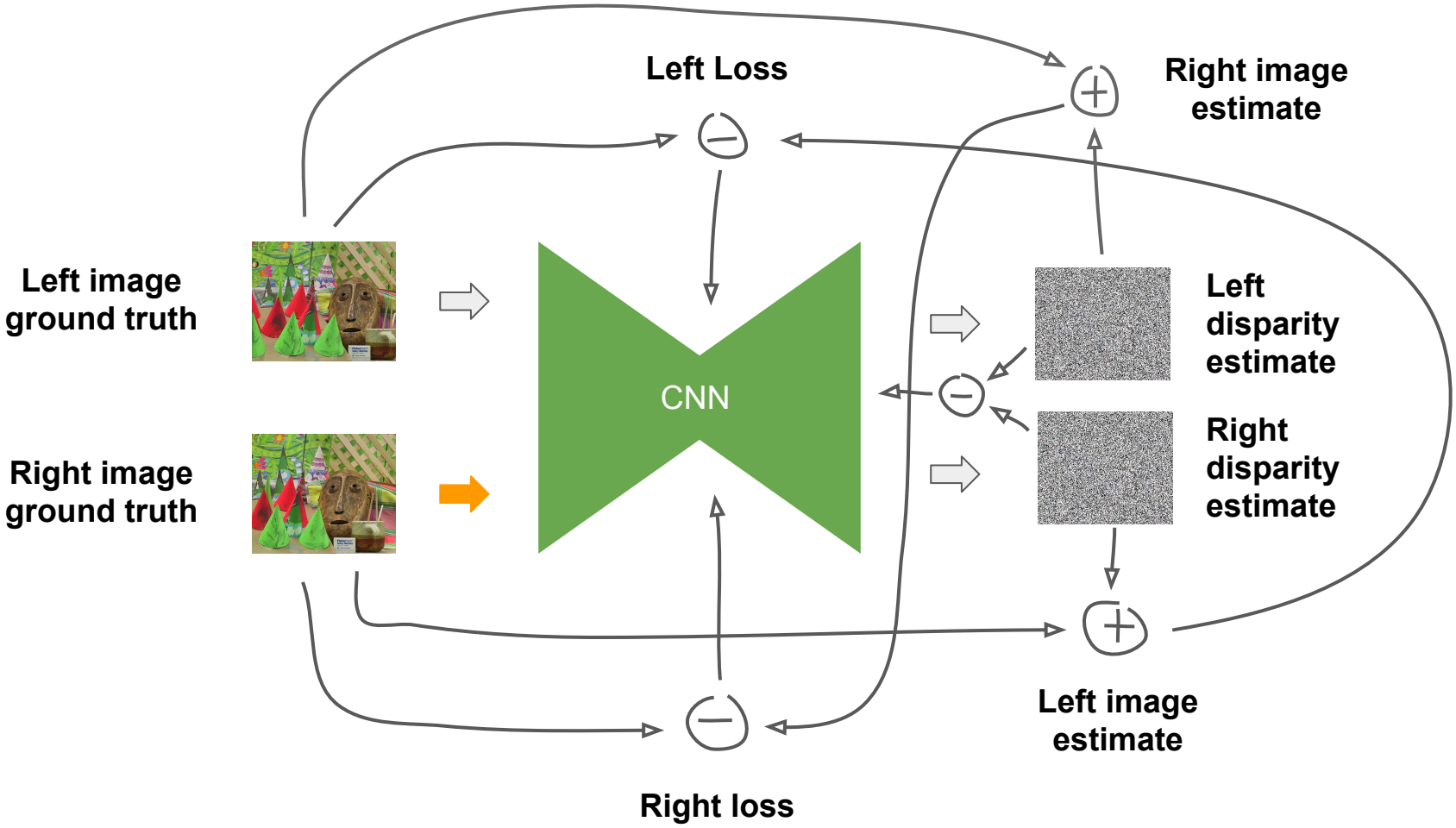

Loss
Left image
ground truth
CNN
Right
disparity
estimate
Left image
estimate

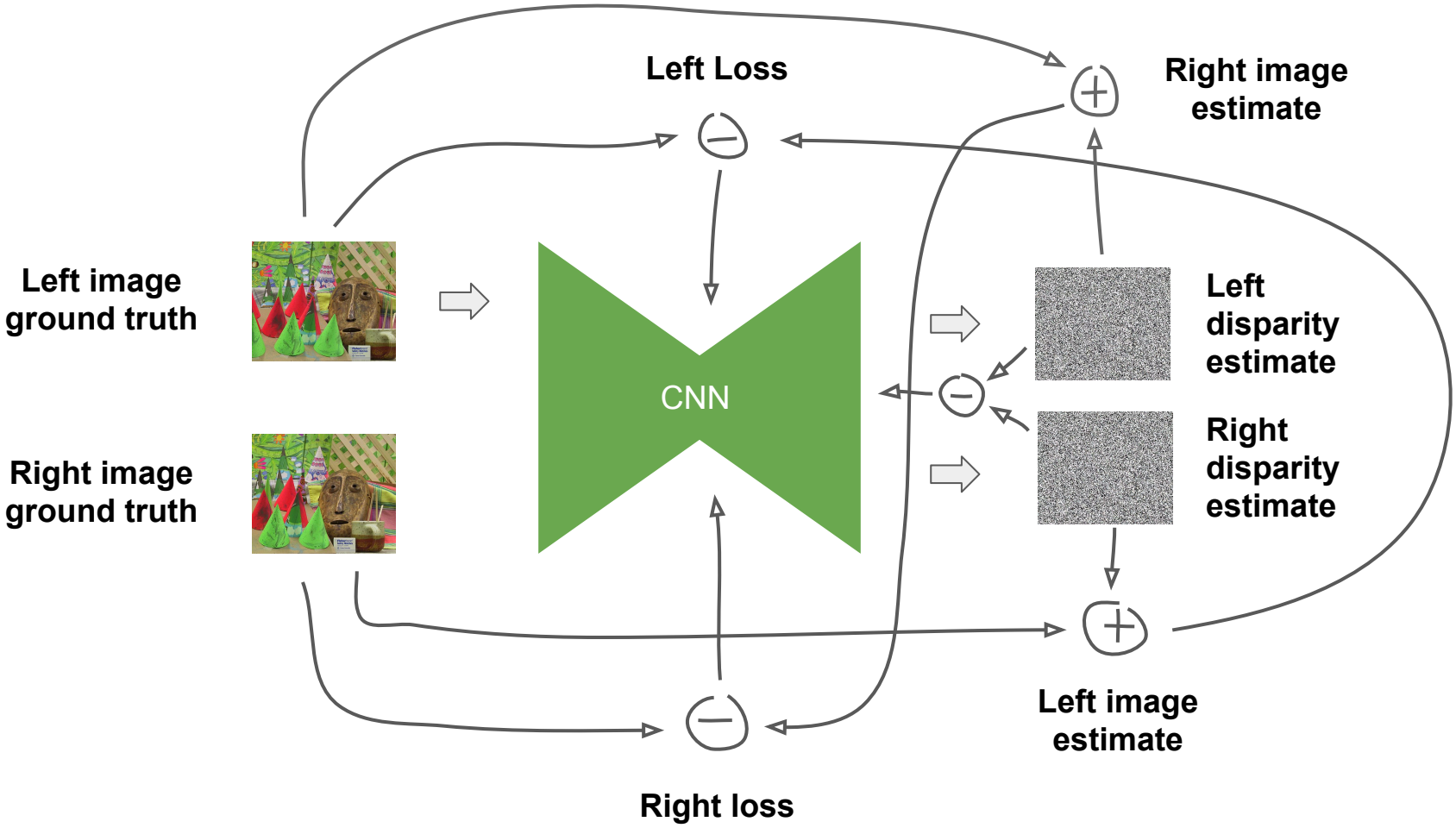

Loss
Left image ground truth
CNN
Left disparity estimate
Right disparity estimate
Left image estimate


Left Loss
Right image estimate
Left image ground truth
CNN
Left disparity estimate
Right image ground truth
Right disparity estimate
Left image estimate
Right loss

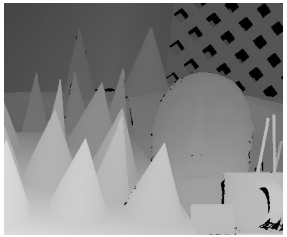

Left Loss
Right image estimate
Left image ground truth
CNN
Left disparity estimate
Right image ground truth
Right disparity estimate
Left image estimate
Right loss


Left Loss
Right image estimate
Left image ground truth
CNN
Left disparity estimate
Right image ground truth
Right disparity estimate
Left image estimate
Right loss


Left Loss
Right image estimate
Left image ground truth
CNN
Left disparity estimate
Right image ground truth
Right disparity estimate
Left image estimate
Right loss


Left Loss
Right image estimate
Left image ground truth
CNN
Left disparity estimate
Right image ground truth
Right disparity estimate
Left image estimate
Right loss

CNN



Unsupervised Monocular Depth Estimation with Left-Right Consistency

Clément Godard Oisin Mac Aodha Gabriel J. Brostow
University College London
http://visual.cs.ucl.ac.uk/pubs/monoDepth/

Abstract

Learning based methods have shown very promising results for the task of depth estimation in single images. However, most existing approaches treat depth prediction as a supervised regression problem and as a result, require vast quantities of corresponding ground truth depth data for training. Just recording quality depth data in a range of environments is a challenging problem. In this paper, we innovate beyond existing approaches, replacing the use of explicit depth data during training with easier-to-obtain binocular stereo footage.

We propose a novel training objective that enables our convolutional neural network to learn to perform single image depth estimation, despite the absence of ground truth depth data. Exploiting epipolar geometry constraints, we generate disparity images by training our network with an image reconstruction loss. We show that solving for image reconstruction alone results in poor quality depth images. To overcome this problem, we propose a novel training loss that enforces consistency between the disparities produced relative to both the left and right

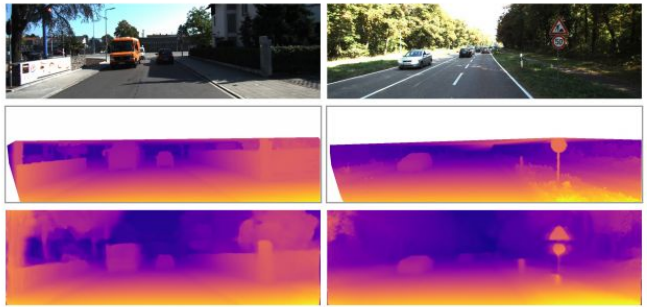


Figure 1. Our depth prediction results on KITTI 2015. Top to bottom: input image, ground truth disparities, and our result. Our method is able to estimate depth for thin structures such as street signs and poles.

have been restricted to scenes where large image collections and their corresponding pixel depths are available.

Understanding the shape of a scene from a single image, independent of its appearance, is a fundamental problem in machine perception. There are many applications such as synthetic object insertion in computer graphics [29], synthetic

Differentiable Bilinear Sampling

Next time