

# Stereo Vision

Jo-won

# 목 차

- Stereo Vision & Image transformation
- Deep learning in Stereo Vision

---

# **Stereo Vision & Image transformation**

---

# 1.1 Homogeneous coordinate

- Scale이 무시되는 좌표계
- Homogeneous 좌표 표현은 무한히 많음

Homogeneous  
좌표 표현

In 2D :  $[x, y] \rightarrow [x, y, 1]$  or  $[wx, wy, w]$

In 3D :  $[x, y, z] \rightarrow [x, y, z, 1]$  or  $[wx, wy, wz, w]$

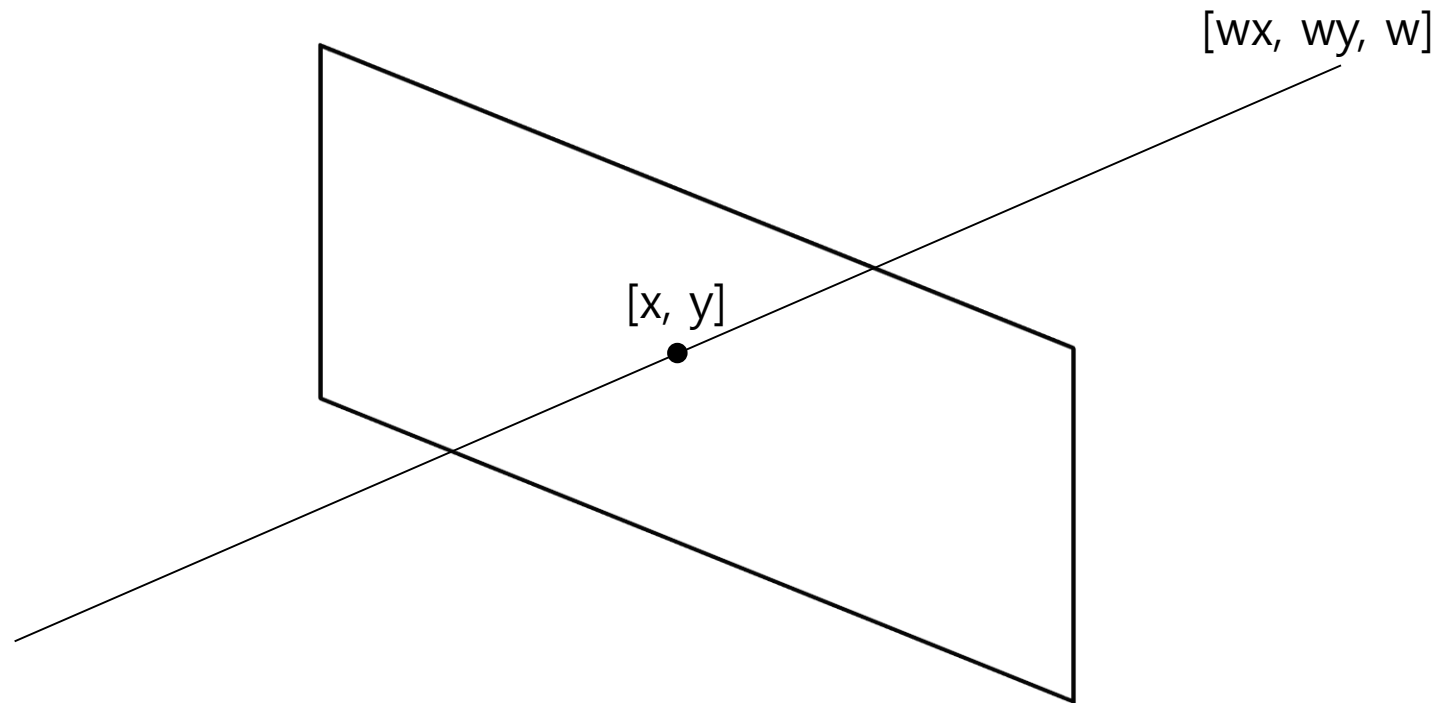
Inverse Homogeneous  
좌표 표현

In 2D :  $[x, y, a] \rightarrow [x/a, y/a, 1]$

In 3D :  $[x, y, z, a] \rightarrow [x/a, y/a, z/a, 1]$   
( $w > 0, a > 0$ )

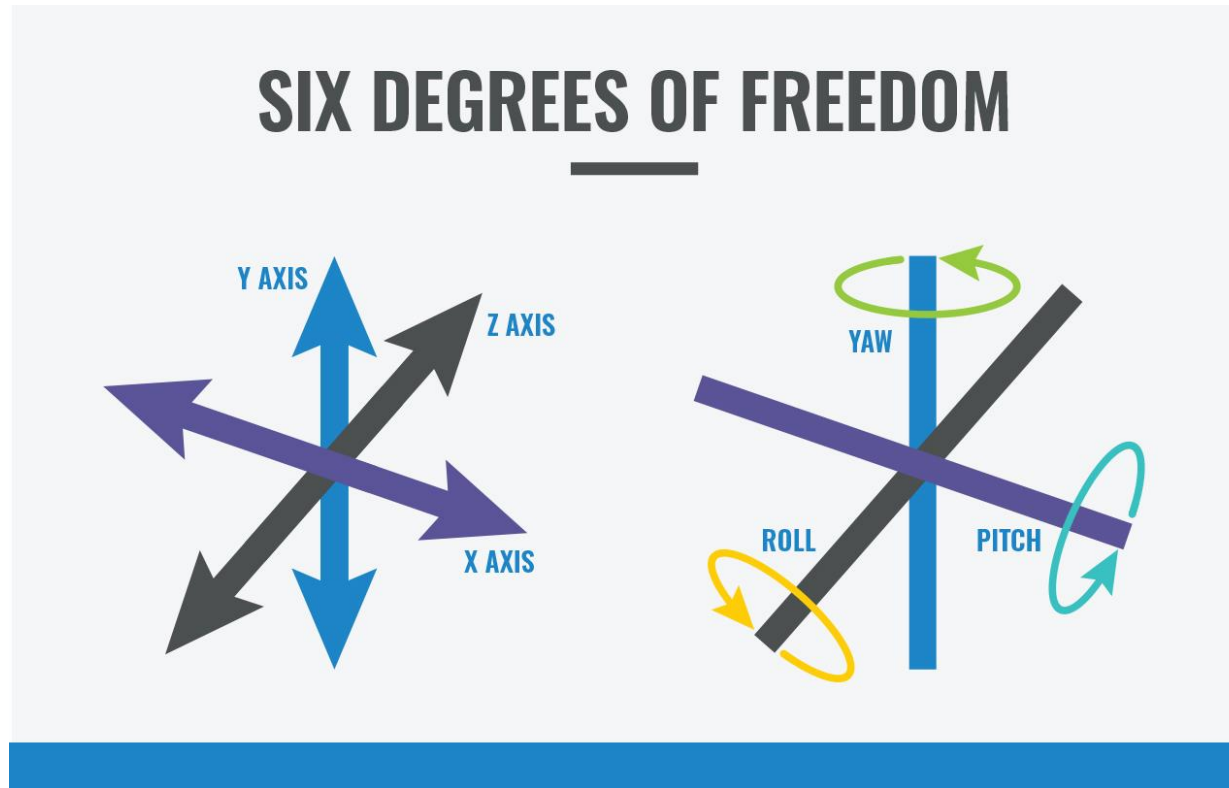
# 1.1 Homogeneous coordinate

- 이미지 평면 상의 한 점의 homogeneous 좌표라는 것은 이 점으로 투영되는 ray상의 모든 점을 한꺼번에 표현하는 방법

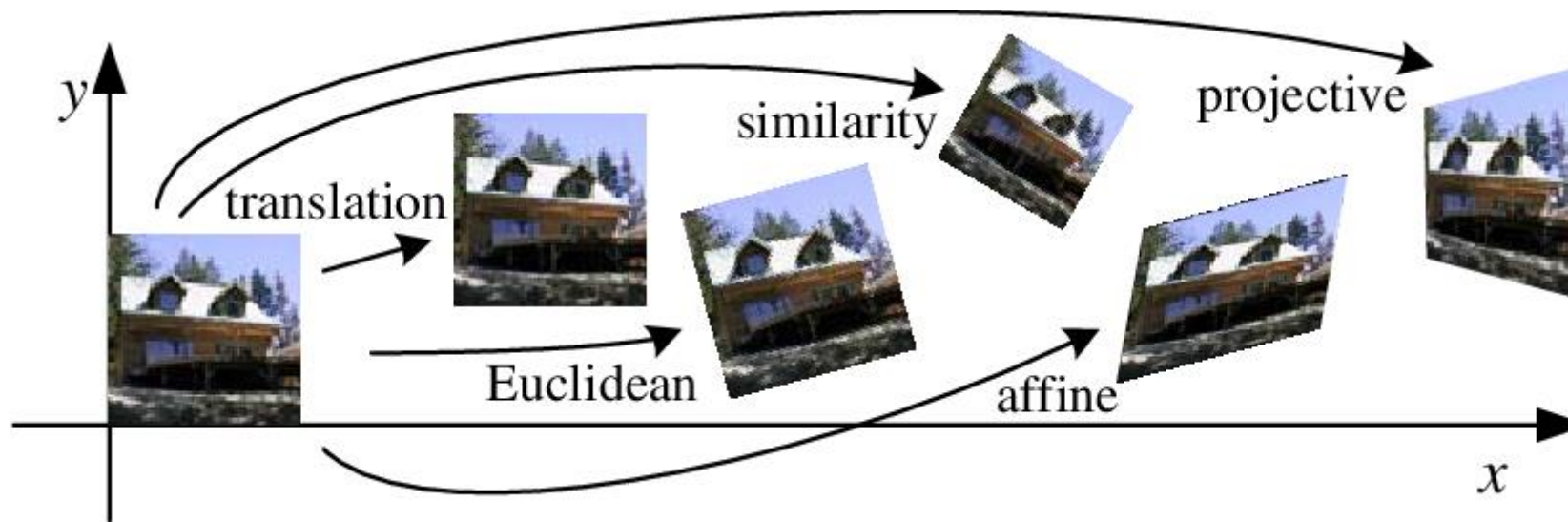


## 1.2 Degree of Freedom (DoF)

- 자유도
- 이미지가 변환될 때 최소한으로 필요한 독립변수의 수



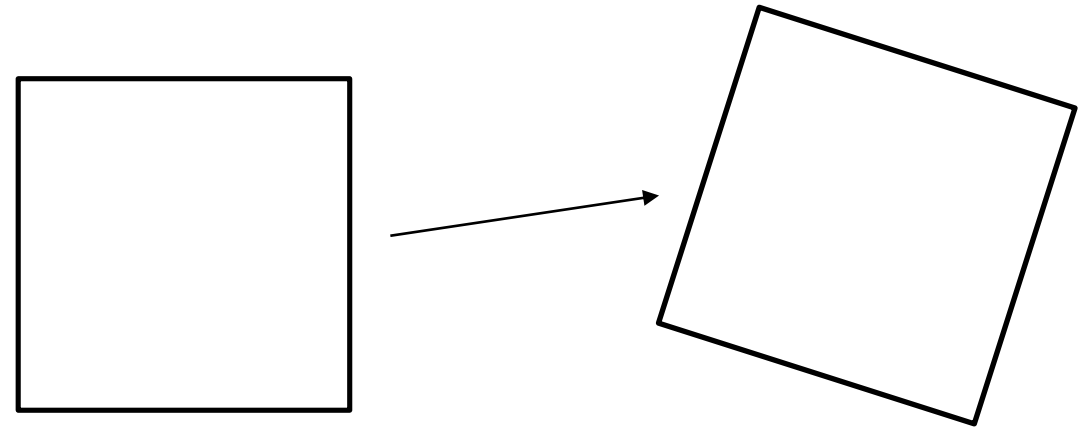
## 1.3 2D Image geometric transformation



# 1.3.1 Rigid transformation (Euclidean transformation)

- 평행이동과 회전이동만을 허용한 변환 (Translation, Rotation)
- 자유도는 3, 변환 특정 시 2쌍의 매칭 쌍 필요

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \underbrace{\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}}_{\text{Rotation}} \begin{bmatrix} x \\ y \end{bmatrix} + \underbrace{\begin{bmatrix} c \\ d \end{bmatrix}}_{\text{Translation}}$$



$$\text{Translation : } \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} x + t_x \\ y + t_y \\ 1 \end{bmatrix}$$

$$\text{Rigid : } \begin{bmatrix} \cos(\theta) & -\sin(\theta) & t_x \\ \sin(\theta) & \cos(\theta) & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix}$$



## 1.3.2 Similarity transformation

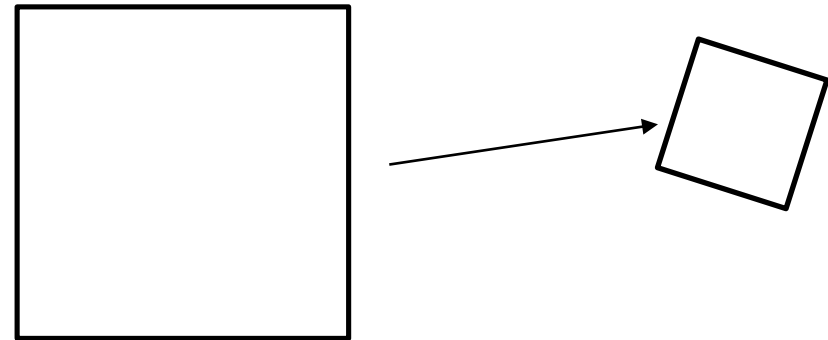
- 평행이동과 회전이동, 크기변화로 구성(Translation, Rotation, Scaling)
- 자유도는 4, 변환 특정 시 2쌍의 매칭 쌍 필요

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \underbrace{\mathbf{S}}_{\text{Scaling}} * \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} c \\ d \end{bmatrix}$$



$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} c \\ d \end{bmatrix}$$

$a = s \cdot \cos \theta, b = s \cdot \sin \theta$



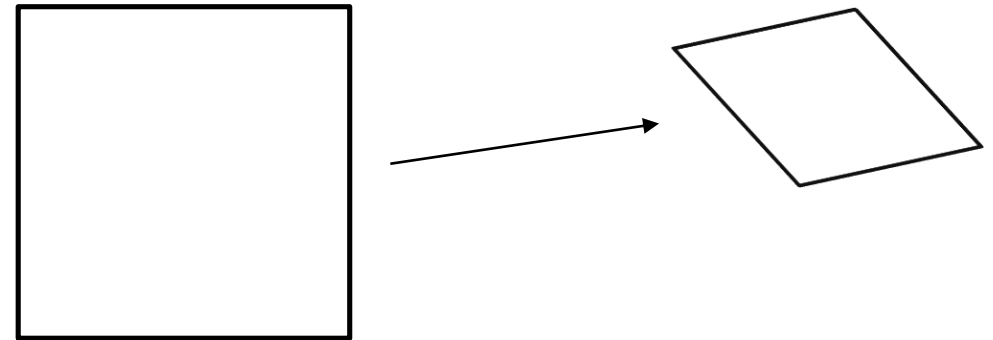
**Similarity :**

$$\begin{bmatrix} s \cos(\theta) & -s \sin(\theta) & t_x \\ s \sin(\theta) & s \cos(\theta) & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix}$$

## 1.3.3 Affine transformation

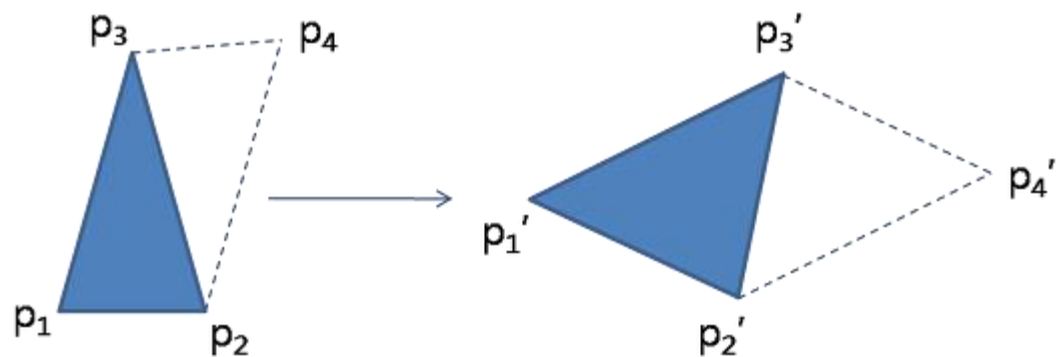
- 평행이동과 회전이동, 크기변화, 전단, 반전으로 구성  
(Translation, Rotation, Scaling, Shearing, Reflect)
- 직선, 길이의 비, 평행성을 유지하는 변환
- 자유도는 6, 변환 특정 시 3쌍의 매칭 쌍 필요

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix}$$



# 1.3.3 Affine transformation

- 임의의 삼각형을 임의의 삼각형으로 평행성을 보존하며 매핑시킬 수 있는 변환



Translation		$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ t_x & t_y & 1 \end{bmatrix}$	$t_x$ specifies the displacement along the $x$ axis $t_y$ specifies the displacement along the $y$ axis.
Scale		$\begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$s_x$ specifies the scale factor along the $x$ axis $s_y$ specifies the scale factor along the $y$ axis.
Shear		$\begin{bmatrix} 1 & sh_y & 0 \\ sh_x & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$sh_x$ specifies the shear factor along the $x$ axis $sh_y$ specifies the shear factor along the $y$ axis.
Rotation		$\begin{bmatrix} \cos(q) & \sin(q) & 0 \\ -\sin(q) & \cos(q) & 0 \\ 0 & 0 & 1 \end{bmatrix}$	$q$ specifies the angle of rotation.

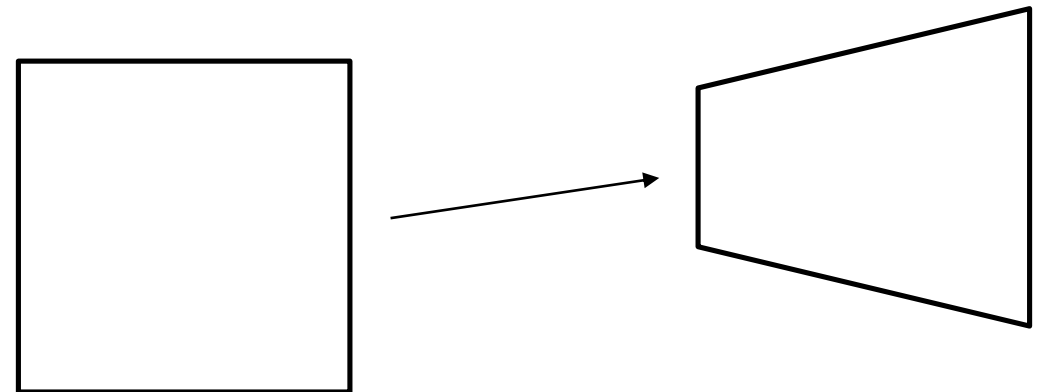
## 1.3.4 Projective transformation (Homography)

- 3D공간 상의 평면 물체를 서로 다른 카메라로 각각 영상 A와 영상 B로 투영되었을 때(perspective transformation) 영상 A와 영상 B의 관계를 homography로 표현 가능
- Homogeneous coordinate에서 정의되어 homography 또한 scale 결정불가
- 자유도 8, 변환 특정 시 4쌍의 매칭 쌍 필요

$$w \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

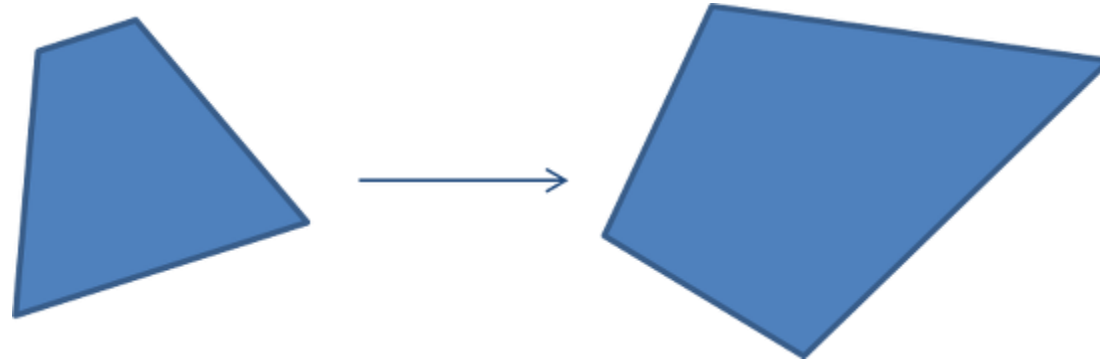
↓  $\times \frac{1}{h_{33}}$  normalize

$$H = \begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & 1 \end{bmatrix}$$




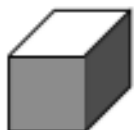


## 1.3.4 Projective transformation

- 2D 평면에서 임의의 사각형을 임의의 사각형으로 매핑 시킬 수 있는 변환



# 1.4 3D transformation

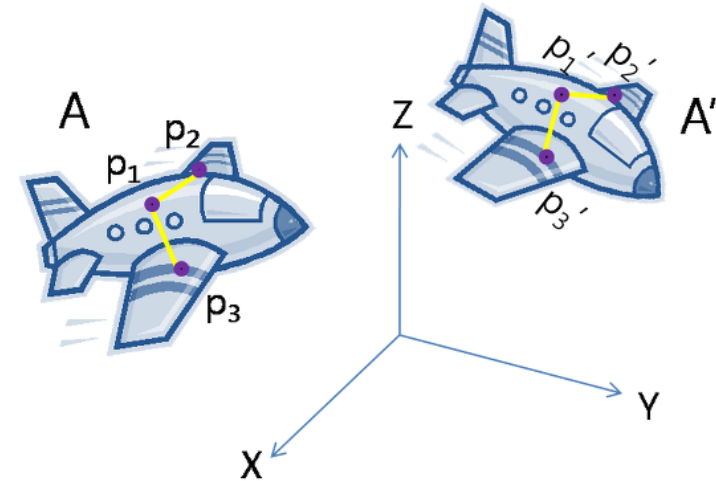
Group	Matrix	Distortion	Invariant properties
Projective 15 dof	$\begin{bmatrix} \mathbf{A} & \mathbf{t} \\ \mathbf{v}^T & v \end{bmatrix}$		Intersection and tangency of surfaces in contact. Sign of Gaussian curvature.
Affine 12 dof	$\begin{bmatrix} \mathbf{A} & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix}$		Parallelism of planes, volume ratios, centroids. The plane at infinity, $\pi_\infty$ , (see section 3.5).
Similarity 7 dof	$\begin{bmatrix} s\mathbf{R} & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix}$		The absolute conic, $\Omega_\infty$ , (see section 3.6).
Euclidean 6 dof	$\begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix}$		Volume.

# 1.4.1 3D Rigid transformation (Euclidean transformation)

- Computer Vision에서 현재 관심있는 3D transformation은 평행이동과 회전이동
- X축, Y축, Z축으로의 회전 각도와 이동거리로 인해 총 자유도는 6
- 변환 특정 시 3쌍의 매칭 쌍 필요

$$\begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix} = R \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}$$

$$\begin{bmatrix} X' \\ Y' \\ Z' \\ 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$



---

# **Deep learning in Stereo Vision**

---



## 2.1 Unsupervised Monocular Depth Estimation with Left-Right Consistency

### Unsupervised Monocular Depth Estimation with Left-Right Consistency

Clément Godard    Oisín 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 images, leading to improved performance and robustness compared to existing approaches. Our method produces state of the art results for monocular depth estimation on the KITTI driving dataset, even outperforming supervised methods that have been trained with ground truth depth.

#### 1. Introduction

Depth estimation from images has a long history in computer vision. Fruitful approaches have relied on structure from motion, shape-from-X, binocular, and multi-view stereo. However, most of these techniques rely on the assumption that multiple observations of the scene of interest are available. These can come in the form of multiple viewpoints, or observations of the scene under different lighting conditions. To overcome this limitation, there has recently been a surge in the number of works that pose the task of monocular depth estimation as a supervised learning problem [32, 10, 36]. These methods attempt to directly predict the depth of each pixel in an image using models that have been trained offline on large collections of ground truth depth data. While these methods have enjoyed great success, to date they

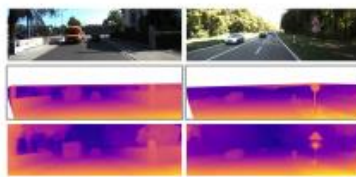


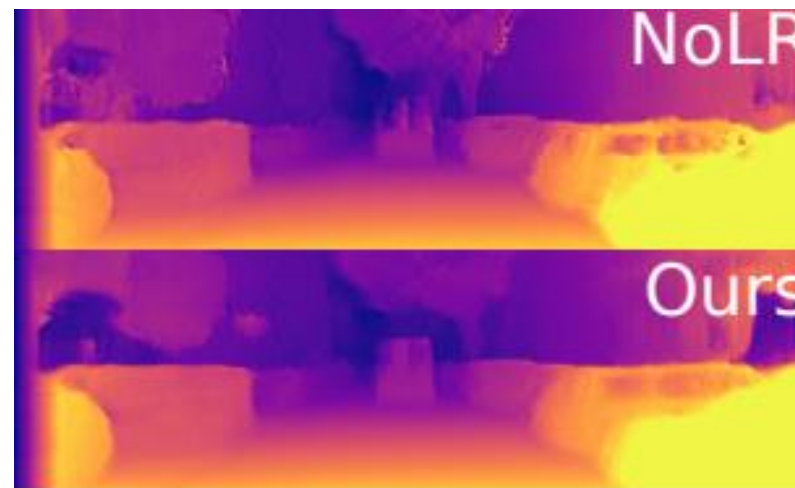
Figure 1. Our depth prediction results on KITTI 2015. Top to bottom: input image, ground truth disparities, and our result. Our method is able to estimate depth for thin structures such as street signs and poles.

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 depth of field in computational photography [3], grasping in robotics [34], using depth as a cue in human body pose estimation [48], robot assisted surgery [49], and automatic 2D to 3D conversion in film [53]. Accurate depth data from one or more cameras is also crucial for self-driving cars, where expensive laser-based systems are often used.

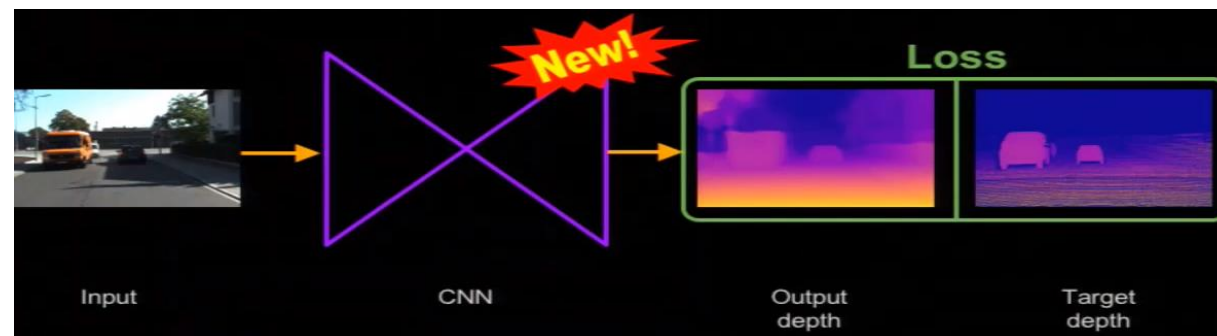
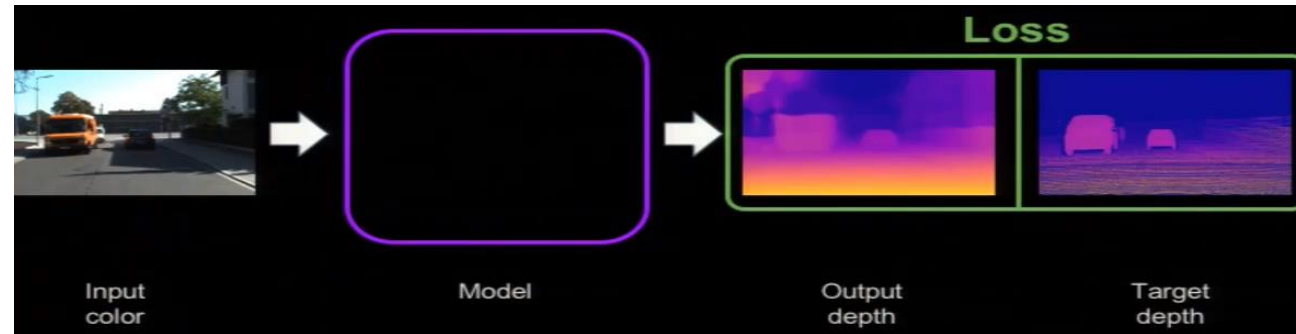
Humans perform well at monocular depth estimation by exploiting cues such as perspective, scaling relative to the known size of familiar objects, appearance in the form of lighting and shading and occlusion [34]. This combination of both top-down and bottom-up cues appears to link full scene understanding with our ability to accurately estimate depth. In this work, we take an alternative approach and treat automatic depth estimation as an image reconstruction problem during training. Our fully convolutional model does not require any depth data, and is instead trained to synthesize depth as an intermediate. It learns to predict the pixel-level correspondence between pairs of rectified stereo images that have a known camera baseline. There are some existing methods that also address the same problem, but with several limitations. For example they are not fully differentiable, making training suboptimal [16], or have image formation models that do

- Train은 stereo calibrated image만을 이용한 unsupervised 방식
- Inference시 한 장의 영상으로 disparity 추출



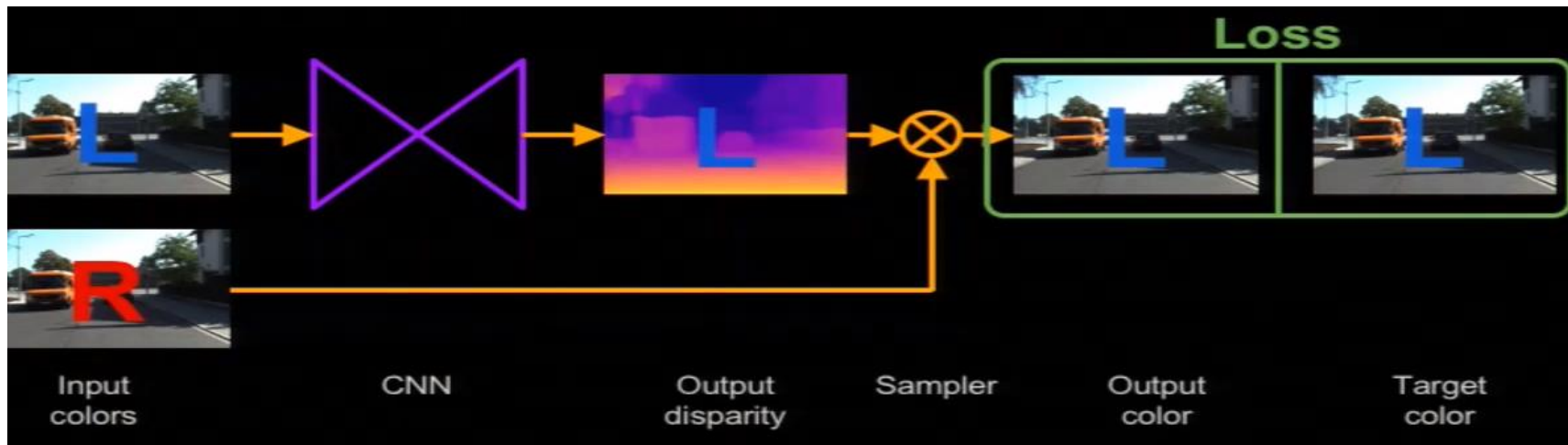
## 2.1.1 Previous approach in depth estimation - Supervised

Naive

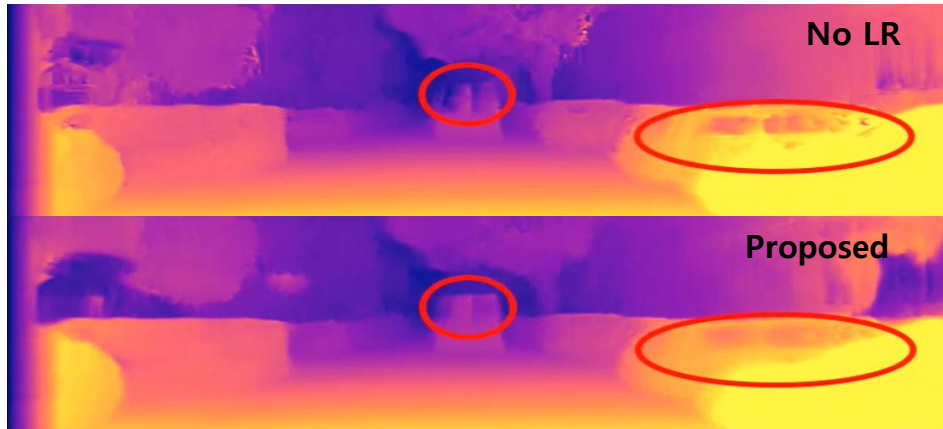


## 2.1.2 Previous approach in depth estimation - Unsupervised

No LR

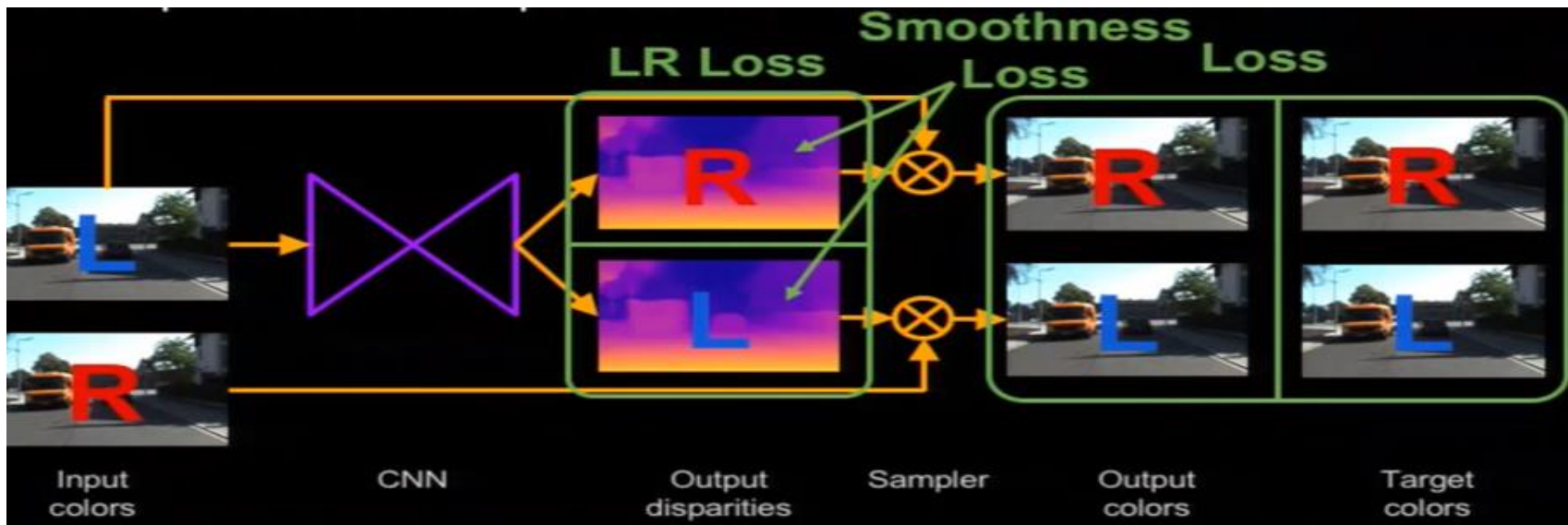


## 2.2 Proposed approach



### Loss

- **Appearance Matching Loss** : disparity를 이용해 반대쪽 영상을 복원한 후 실제 영상과의 차이, SSIM을 이용
- **Disparity smoothness Loss** : 더 정확하게 하고 visual artifact들을 줄이기 위해 edge에 낮은 가중치
- **Left-Right Disparity Consistency Loss** : disparity가 정확할수록 left-to-right disparity와 right-to-left disparity는 역관계가 되어야함



### SSIM

- 구조적 유사지수
- 압축 및 변환에 의해 발생하는 왜곡에 대하여 원본 영상의 유사도를 측정하는 방법

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(2\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

- $\mu_x$  the average of  $x$ ;
- $\mu_y$  the average of  $y$ ;
- $\sigma_x^2$  the variance of  $x$ ;
- $\sigma_y^2$  the variance of  $y$ ;
- $\sigma_{xy}$  the covariance of  $x$  and  $y$ ;
- $c_1 = (k_1 L)^2$ ,  $c_2 = (k_2 L)^2$  two variables to stabilize the division with weak denominator;
- $L$  the dynamic range of the pixel-values (typically this is  $2^{\text{#bits per pixel}} - 1$ );
- $k_1 = 0.01$  and  $k_2 = 0.03$  by default.



## 2.2 GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose



This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.  
Except for this watermark, it is identical to the version available on IEEE Xplore.

### GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose

Zhichao Yin and Jianping Shi  
SenseTime Research

{yinzichao, shijianping}@sensetime.com

#### Abstract

We propose GeoNet, a jointly unsupervised learning framework for monocular depth, optical flow and ego-motion estimation from videos. The three components are coupled by the nature of 3D scene geometry, jointly learned by our framework in an end-to-end manner. Specifically, geometric relationships are extracted over the predictions of individual modules and then combined as an image reconstruction loss, reasoning about static and dynamic scene parts separately. Furthermore, we propose an adaptive geometric consistency loss to increase robustness towards outliers and non-Lambertian regions, which resolves occlusions and texture ambiguities effectively. Experimentation on the KITTI driving dataset reveals that our scheme achieves state-of-the-art results in all of the three tasks, performing better than previously unsupervised methods and comparably with supervised ones.

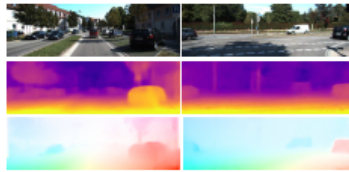


Figure 1. Example predictions by our method on KITTI 2015 [31]. Top to bottom: input image (one of the sequence), depth map and optical flow. Our model is fully unsupervised and can handle dynamic objects and occlusions explicitly.

major advantage comes from big data, which helps capturing high-level semantic correspondences for low level clue learning, thus performing better even in ill-posed regions compared with traditional methods.

Nevertheless, to preserve high performance with more general scenarios, large corpus of groundtruth data are usually needed for deep learning. In most circumstances, expensive laser-based setups and differential GPS are required, restricting the data grow to a large scale. Moreover, previous deep models are mostly tailored to solve one specific task, such as depth [26], optical flow [8], camera pose [22], etc. They do not explore the inherent redundancy among these tasks, which can be formulated by geometry regularities via the nature of 3D scene construction.

Recent works have emerged to formulate these problems together with deep learning. But all possess certain inherent limitations. For example, they require large quantities of laser scanned depth data for supervision [48], demand stereo cameras as additional equipment for data acquisition [15], or cannot explicitly handle non-rigidity and occlusions [50, 56].

In this paper, we propose an unsupervised learning framework GeoNet for jointly estimating monocular depth, optical flow and camera motion from video. The foundation of our approach is built upon the nature of 3D scene geometry (see Sec. 3.1 for details). An intuitive explanation is that most of the natural scenes are comprised of rigid static sur-

- Depth estimation, Visual odometry, Optical flow에 대해 각각 subnetwork를 두고 모두 추론 하는 모델
- DepthNet과 PoseNet을 먼저 학습시킨 후 고정한 뒤 ResFlowNet 학습
- Forward와 backward 두 방향의 flow를 구함

