# Interpretability of CNN

20200827_文献分享_PANZHIYU

2020/9/4

# What is interpretability

- Simulatability (可仿真性): 对整体模型的可理解性。一个模型越简单，它的可理解性就越高。例如：线性分类器或者回归器就是完全可理解的。
- Decomposability (可分解性)： Modularized analysis: the inner working of a complicated system is factorized as a combination of functionalized modules.
- Algorithmic Transparency (算法透明度)： Understand the train algorithm. 网络的目标函数优化往往是非凸的，训练得不到一个唯一解。但SGD-based的方法训练的结果往往都表现不错。

- In summary: Understanding the model; Understanding the train.

# Why interpretability is difficult?

- Human: Even the most talented physicists know little about the essence of this problem, let alone to fully understand the predications of the neural network.
- Cost: …
- Data: 数据源复杂 (it is often very hard to have high quality data such as structured data in many domains); 理解高维数据的mapping 复杂
- Algorithm: Compared to classical convex optimization problems, optimizing a deep learning model is a complex non-convex optimization problem, which is rather hard to comprehend.

- Network Dissection: Quantifying Interpretability of Deep Visual Representations  CVPR 2017
- Interpretable Convolutional Neural Networks CVPR 2018
- Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention ICCV 2017
- Interpretable Transformations with Encoder-Decoder Networks ICCV 2017
- Lightweight Multi-View 3D Pose Estimation through Camera-Disentangled Representation CVPR 2020

# Network Dissection: Quantifying Interpretability of Deep Visual Representations

David Bau∗, Bolei Zhou∗, Aditya Khosla, Aude Oliva, and Antonio Torralba

CSAIL, MIT

CVPR 2017

# Target

- 以往文献说明网络的可解释性通过可视化，没有一个量化的衡量，本文采用分析性框架量化可解释性 （Metric: part interpretability）

→Define it in terms of alignment with a set of human-interpretable concepts

Representation disentangled, single units and single interpretable concepts

计算具有解释性的unit个数，以及具有单一概念的可解释性的个数

# Broden dataset

- a ground truth set of exemplars for a broad set of visual concepts


Figure 2. Samples from the **Broden** Dataset. The ground truth for each concept is a pixel-wise dense annotation.

Table 1. Statistics of each label type included in the data set.

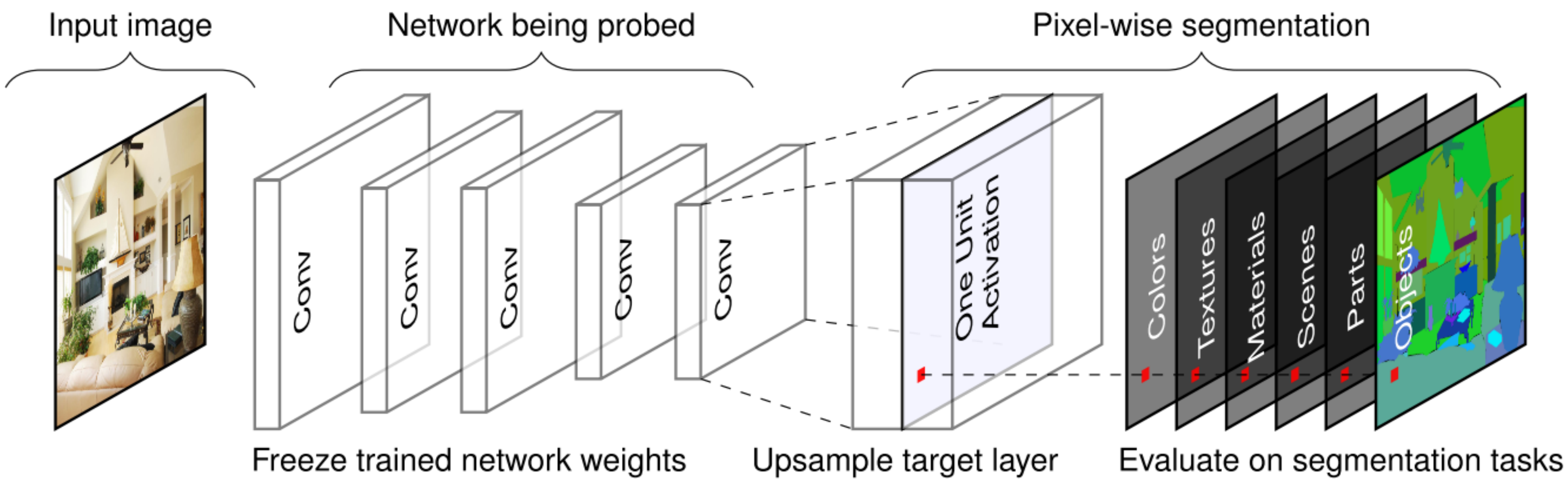
| Category | Classes | Sources | Avg sample |
|---|---|---|---|
| scene | 468 | ADE [43] | 38 |
| object | 584 | ADE [43], Pascal-Context [19] | 491 |
| part | 234 | ADE [43], Pascal-Part [6] | 854 |
| material | 32 | OpenSurfaces [4] | 1,703 |
| texture | 47 | DTD [7] | 140 |
| color | 11 | Generated | 59,250 |

# Scoring Unit Interpretability

- 1. 得到图片x 关于滤波器f的激活结果$A_k(x)$, 然后得到激活结果的分布(distribution) ak.
- 2. 确定top quantile level Tk ,选择原则 $P(a_k > T_k) = 0.005$
- 3. 对$A_k(x)$进行上采样到原图大小(双线性插值)，$M_k(\mathbf{x}) \equiv S_k(\mathbf{x}) \geq T_k$

$$IoU_{k,c} = \frac{\sum |M_k(\mathbf{x}) \cap L_c(\mathbf{x})|}{\sum |M_k(\mathbf{x}) \cup L_c(\mathbf{x})|}$$

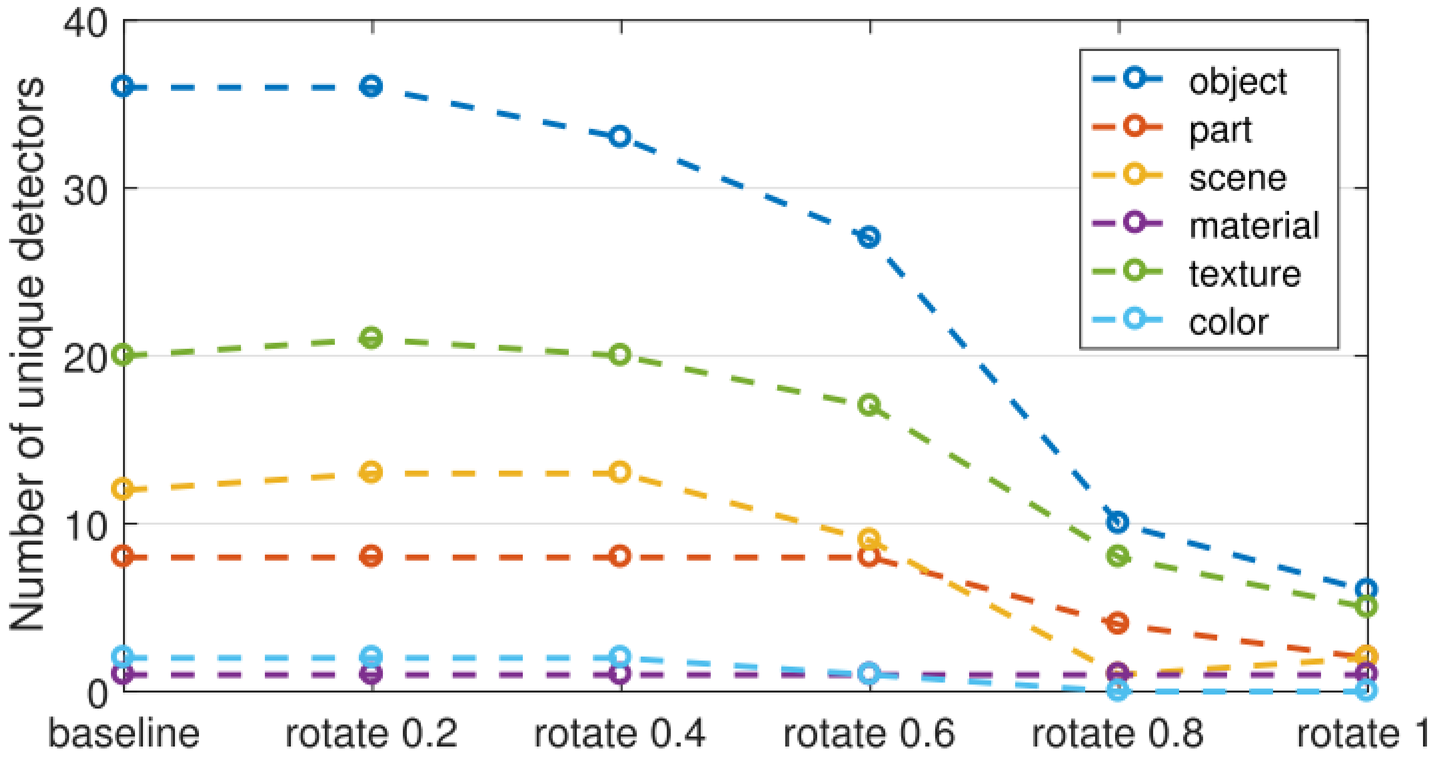
$$IoU_{k,c} > 0.04$$



# Result & Conclusion

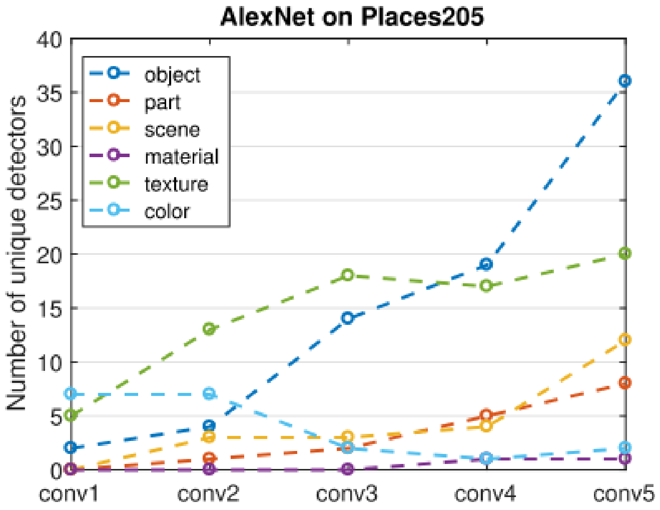
| | conv1 | conv2 | conv3 | conv4 | conv5 |
|---|---|---|---|---|---|
| Interpretable units | 57/96 | 126/256 | 247/384 | 258/384 | 194/256 |
| Human consistency | 82% | 76% | 83% | 82% | 91% |
| Network Dissection | 37% | 56% | 54% | 59% | 71% |

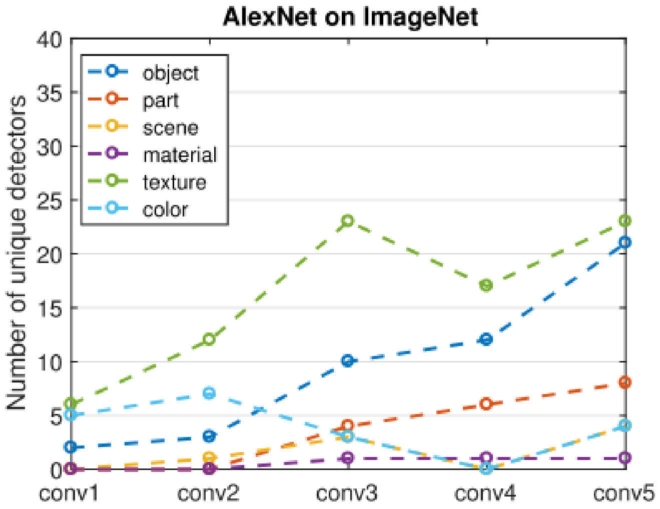
Human evaluation

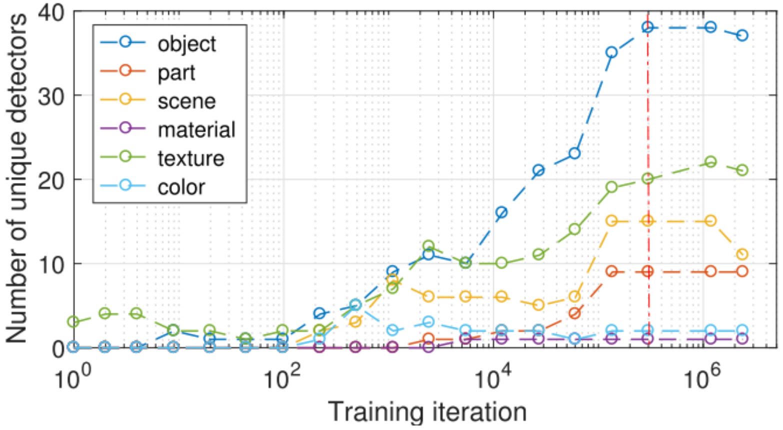


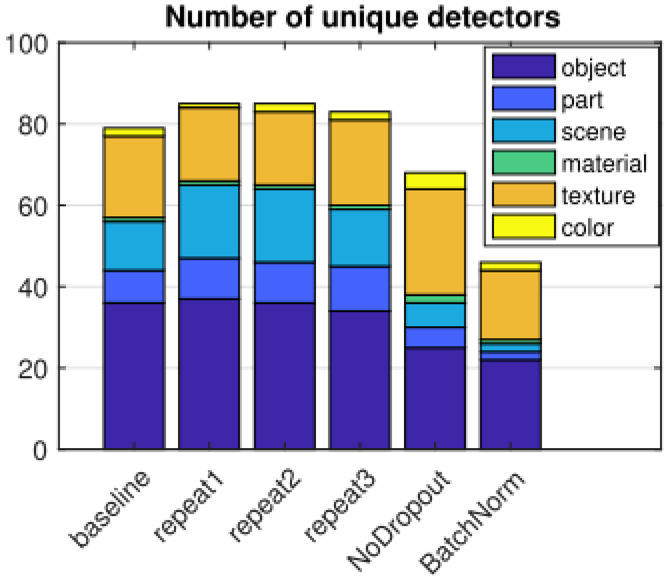

Axis-Aligned Interpretability

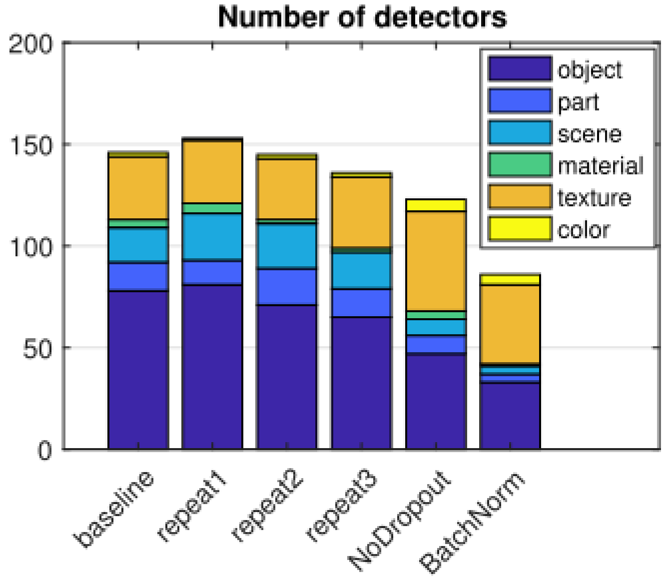
# Result & Conclusion

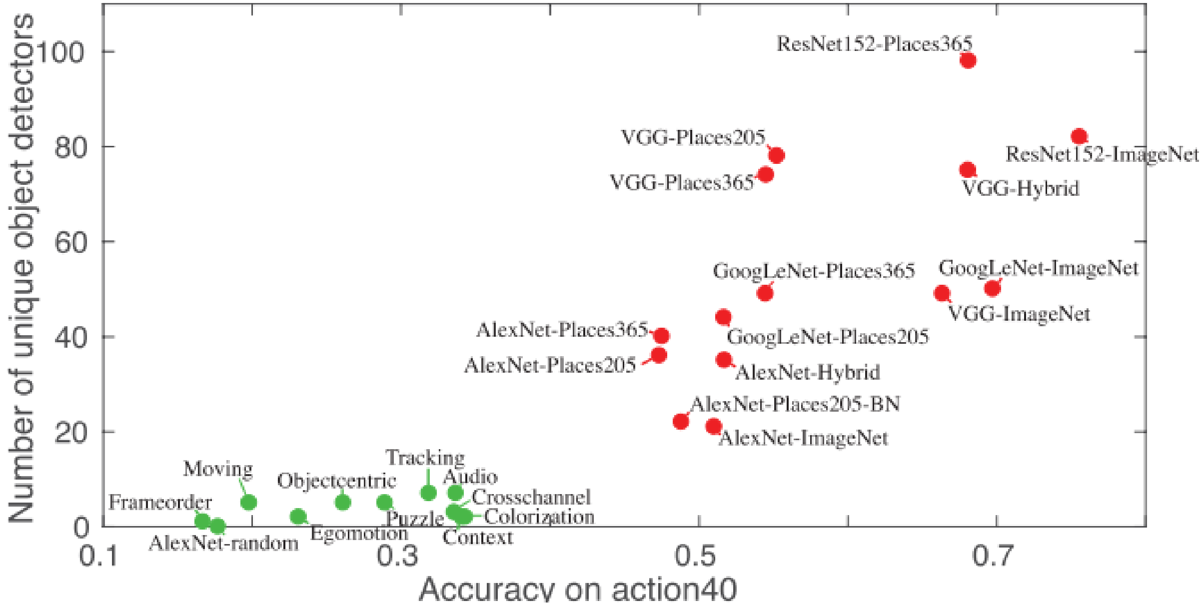


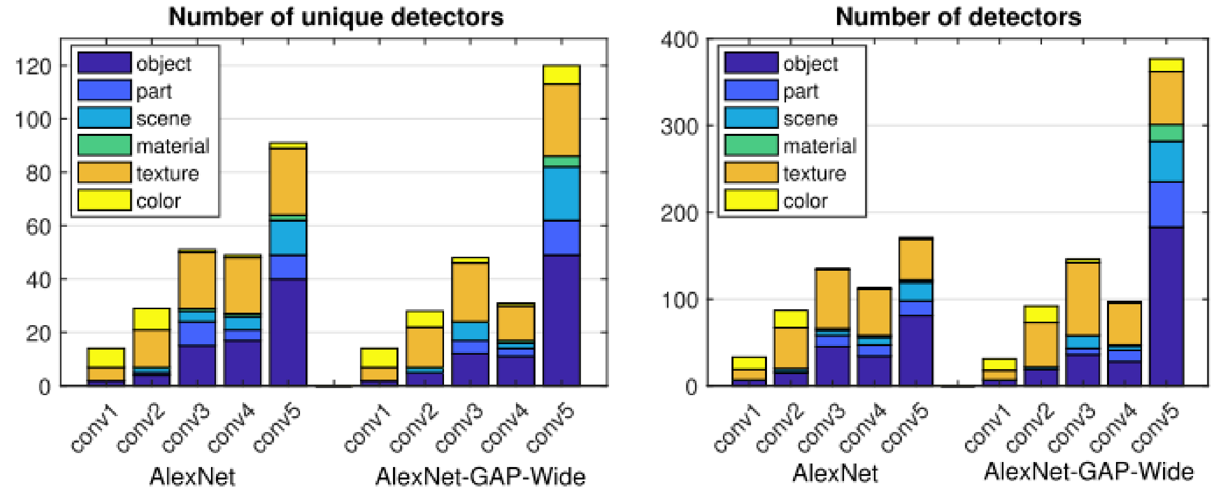






Accuracy on a representation when applied to a task is dependent not only on the number of concept detectors in the representation, but on the suitability of the set of represented concepts to the transfer task




宽度对可解释性有促进作用，但有limitation

# Interpretable Convolutional Neural Networks

Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu

University of California, Los Angeles

CVPR 2018

# Target

让网络的高层滤波器关注一个类别物品的一个局部 （an object part）

Filters in an interpretable CNN are more semantically meaningful than those in traditional CNNs

我们可以更清楚的看到网络记住了哪些特征，更容易相信网络的预测结果

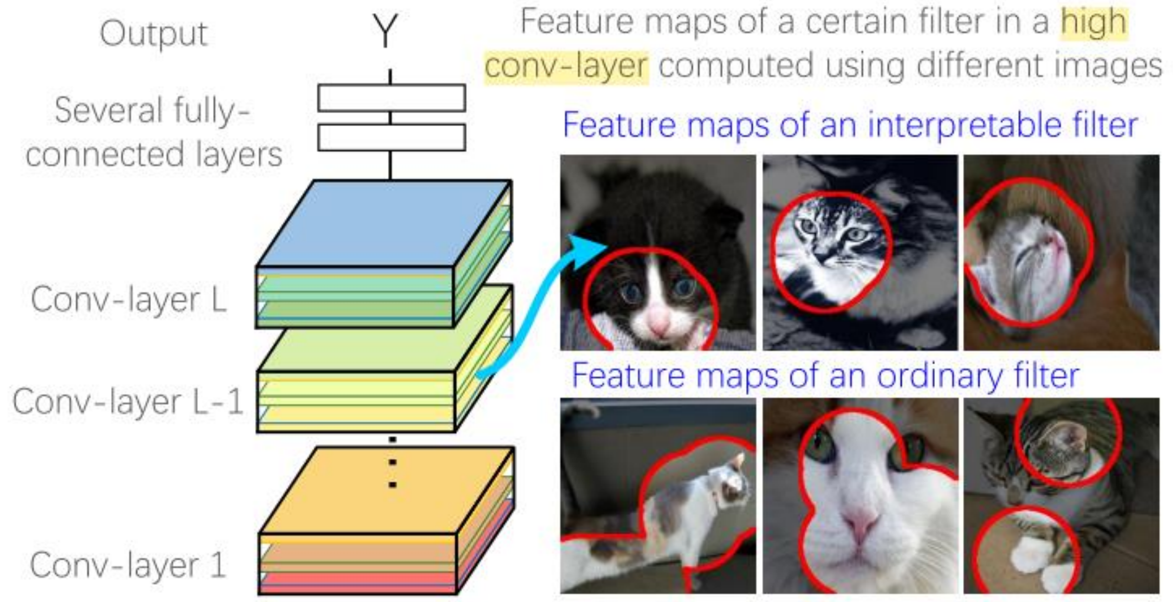
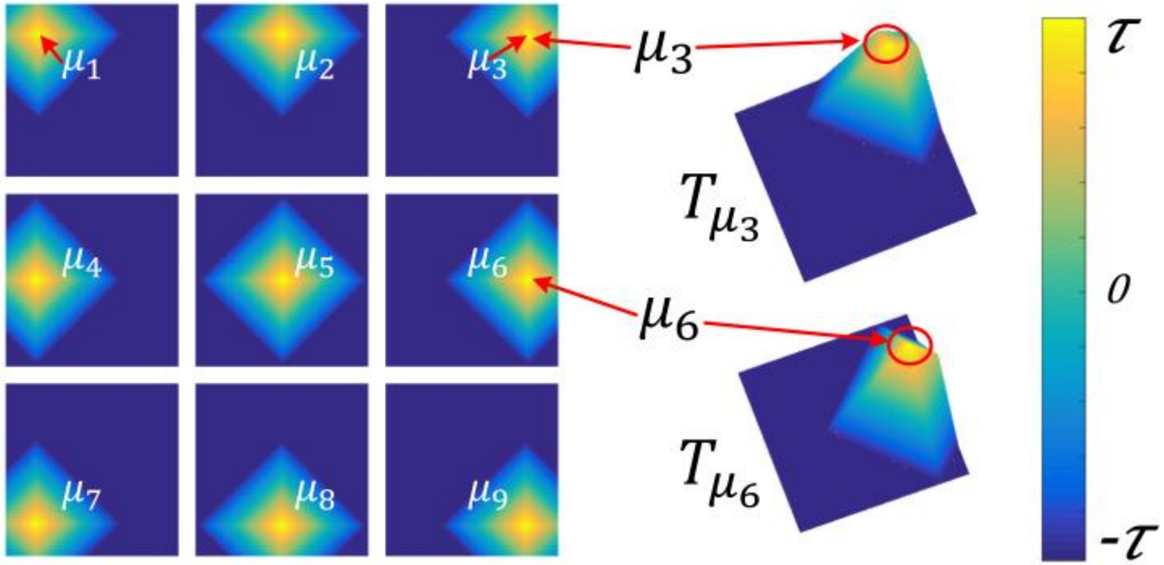
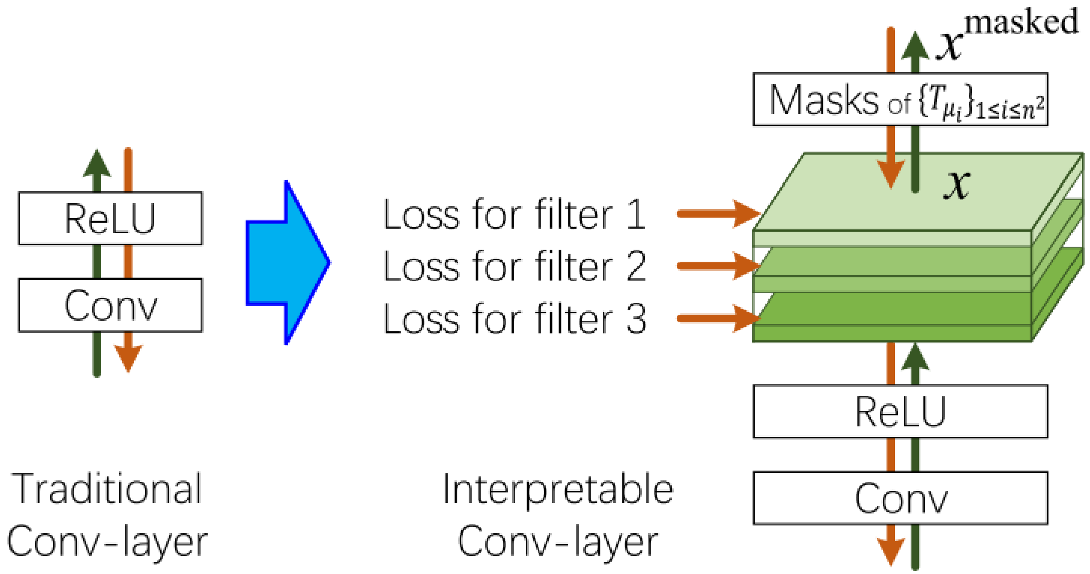



Figure 1. Comparison of a filter's feature maps in an interpretable CNN and those in a traditional CNN.

# Method

- 整体思路： template matching
- Template: 正向传播时作为mask (filter out the noisy activations)， 反向传播用于跟特征图匹配算Loss



# Back prop & Loss

$$\mathbf{T} = \{T^-, \mathbf{T}^+\}$$

- if I ∈ Ic, the feature map x is expected to the assigned template Tˆμ, if I ~∈ Ic we design a negative template T− and hope the feature map x matches to T−.

$$\begin{aligned}\mathbf{Loss}_f &= -MI(\mathbf{X};\mathbf{T}) \quad \text{for filter } f \\ &= -\sum_T p(T)\sum_x p(x|T)\log\frac{p(x|T)}{p(x)}\end{aligned}$$

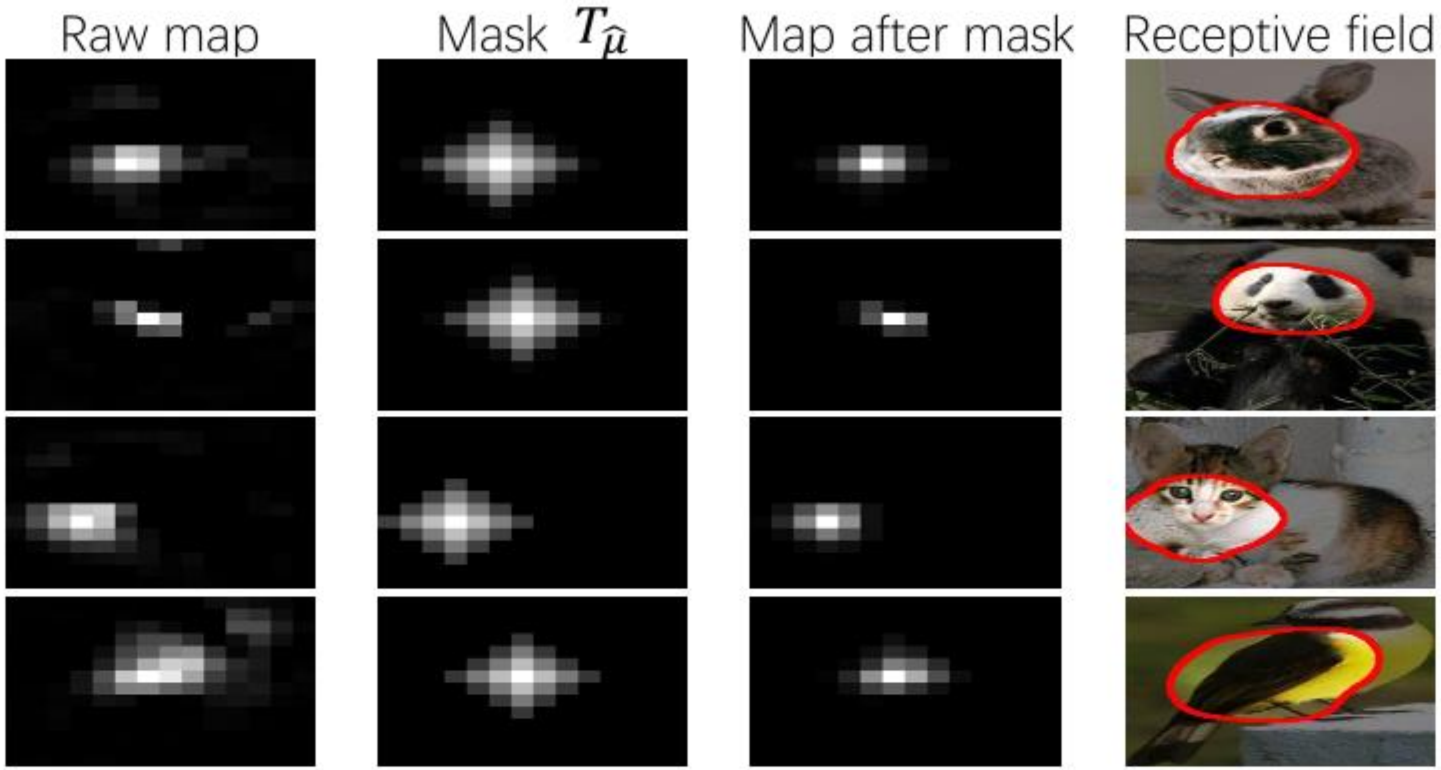
$$\begin{aligned}\mathbf{Loss}_f = &-H(\mathbf{T}) + H(\mathbf{T}' = \{T^-, \mathbf{T}^+\}|\mathbf{X}) \\ &+ \sum_x p(\mathbf{T}^+, x) H(\mathbf{T}^+|X = x)\end{aligned}$$

$$H(\mathbf{T}' = \{T^-, \mathbf{T}^+\}|\mathbf{X}) = -\sum_x p(x) \sum_{T \in \{T^-, \mathbf{T}^+\}} p(T|x)\log p(T|x)$$

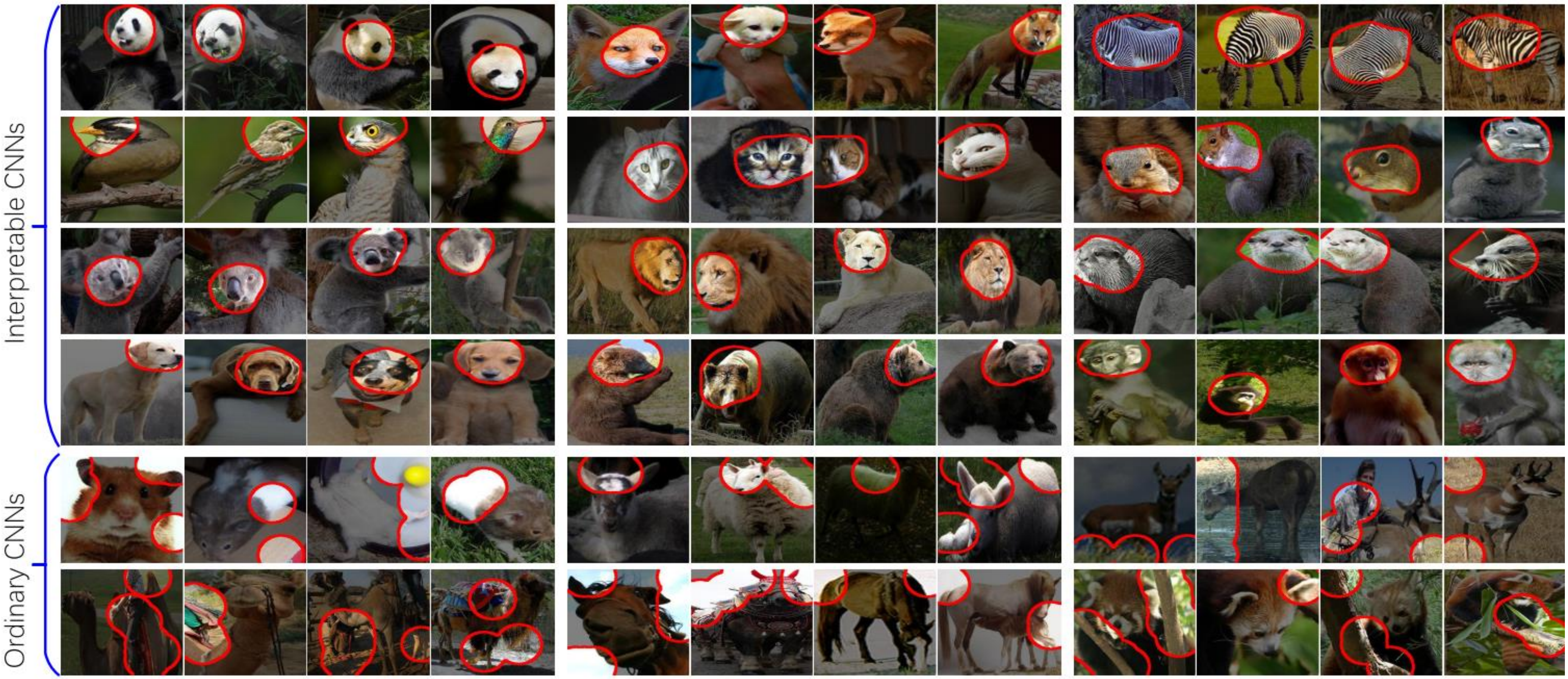
Low inter-category entropy ; one single category

$$H(\mathbf{T}^+|X = x) = \sum_\mu \tilde{p}(T_\mu|x)\log\tilde{p}(T_\mu|x)$$

Low spatial entropy; one single region


Raw map
Mask $T_{\hat{\mu}}$
Map after mask
Receptive field

# Result



# Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention


Jinkyu Kim and John Canny

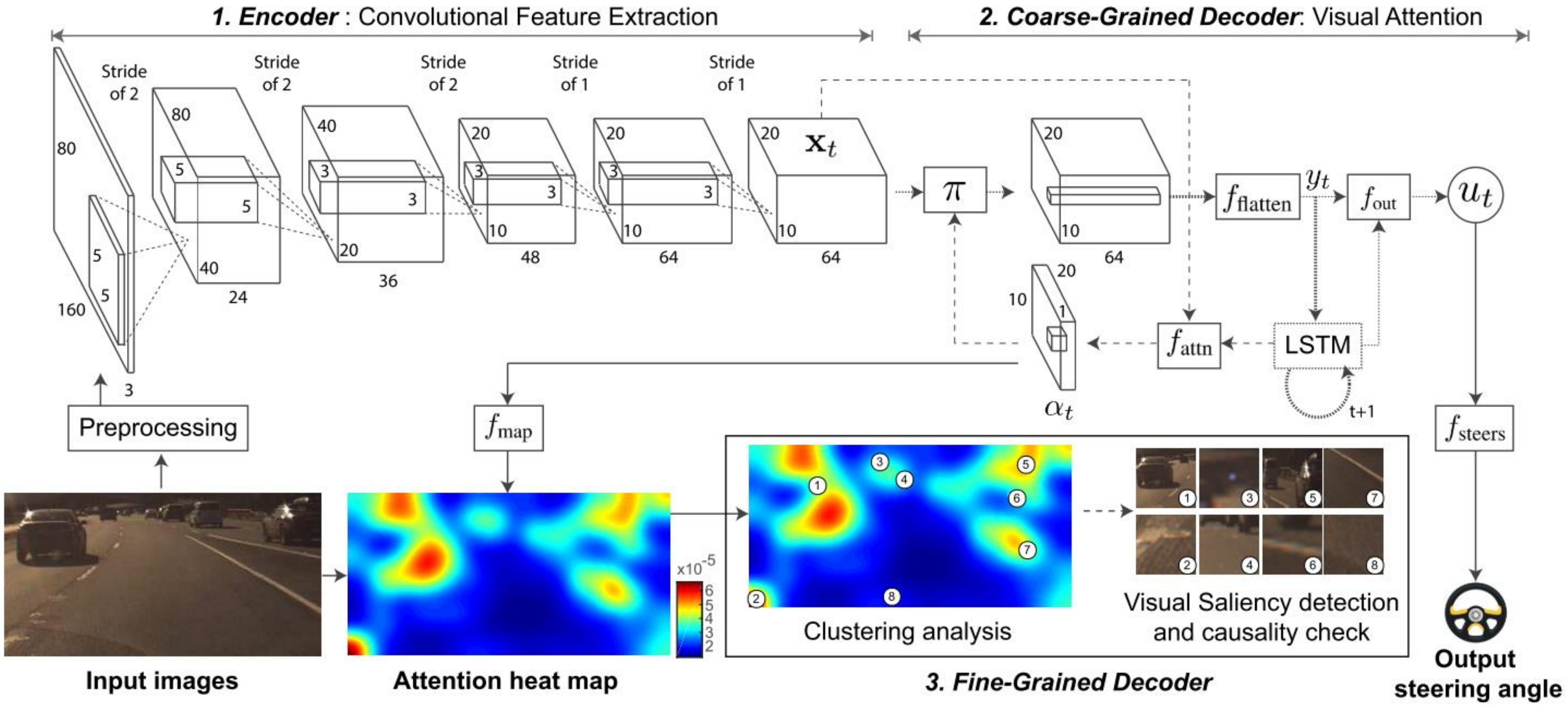
EECS, UC Berkeley, Berkeley, CA 94709, USA



ICCV 2017

# Target

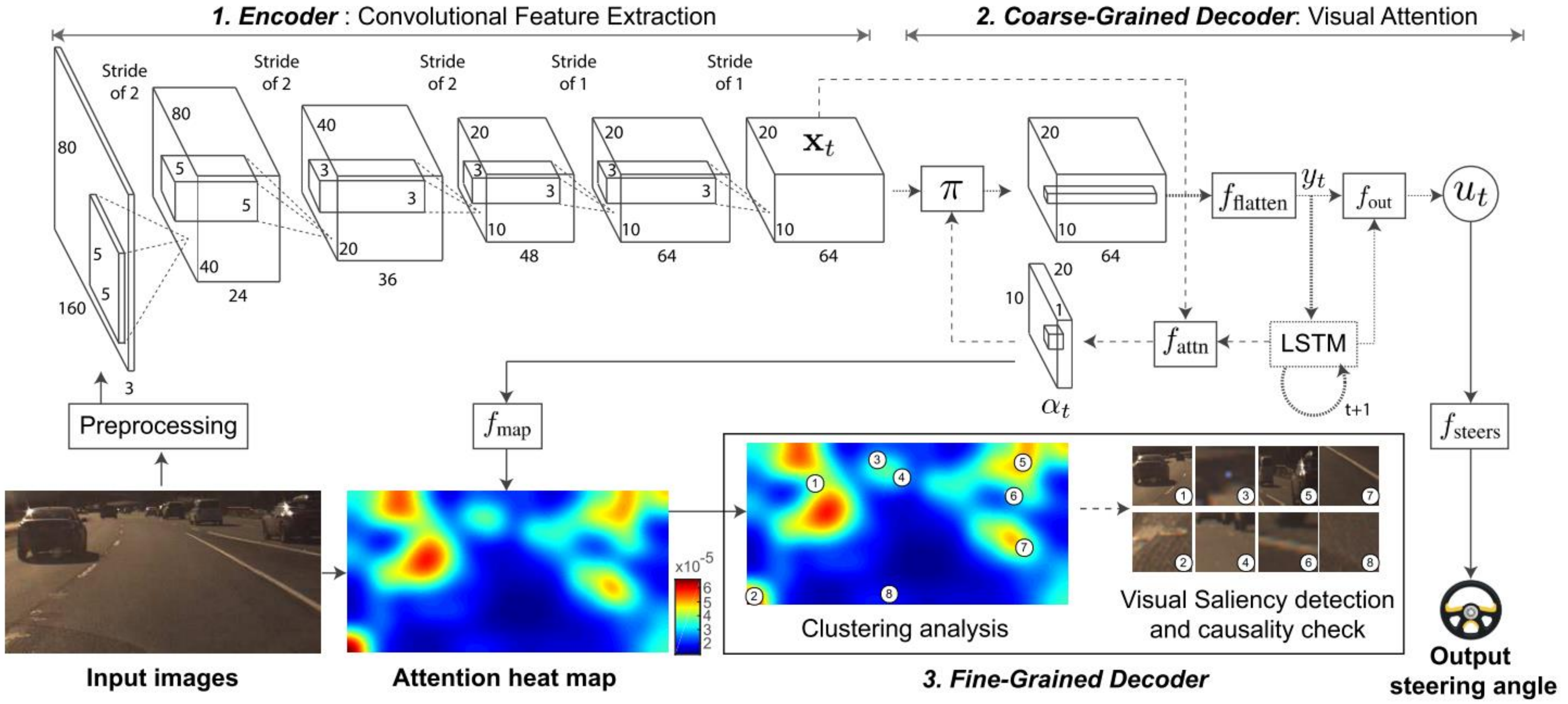
- 网络为E2E实现预测无人驾驶的操作角度(steer angle)
- 同时网络输出attention map， → 通过注意场景的哪部分做出的判断

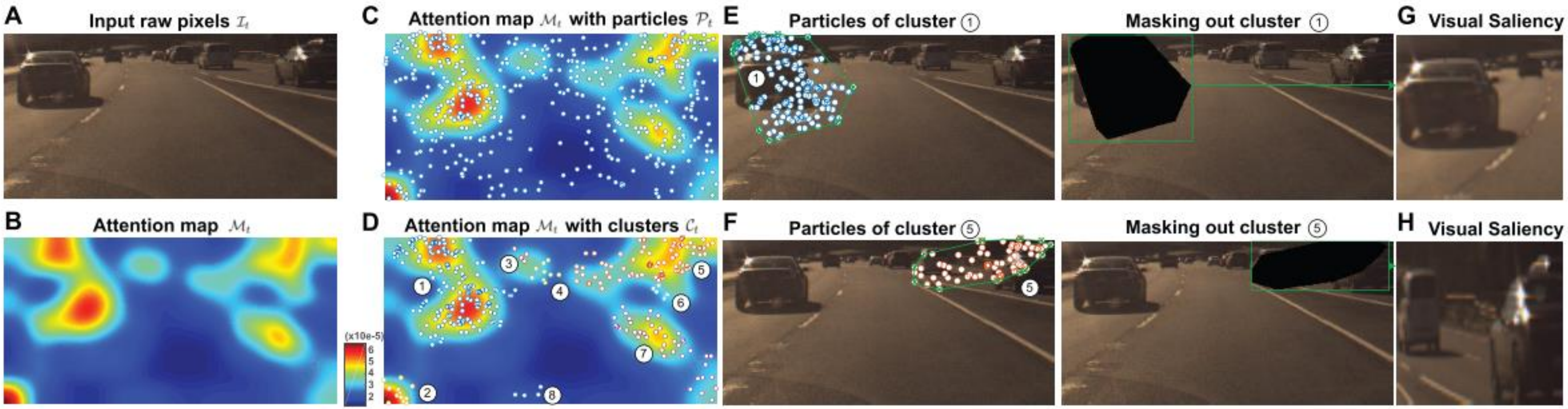


# Network

$$\alpha_{t,i} = \frac{\exp(f_{\text{attn}}(x_{t,i}, h_{t-1}))}{\sum_{j=1}^{L} \exp(f_{\text{attn}}(x_{t,j}, h_{t-1}))}$$

$$\mathcal{L}_1(u_t, \hat{u}_t) = \sum_{t=1}^{T} |u_t - \hat{u}_t| + \lambda \sum_{i=1}^{L} \left(1 - \sum_{t=1}^{T} \alpha_{t,i}\right)$$





A
Input raw pixels $\mathcal{I}_t$
B
Attention map $\mathcal{M}_t$
(x10e-5)
C
Attention map $\mathcal{M}_t$ with particles $\mathcal{P}_t$
D
Attention map $\mathcal{M}_t$ with clusters $\mathcal{C}_t$
E
Particles of cluster 1
Masking out cluster 1
F
Particles of cluster 5
Masking out cluster 5
G
Visual Saliency
H
Visual Saliency

# Interpretable Transformations with Encoder-Decoder Networks

Daniel E. Worrall Stephan J. Garbin Daniyar Turmukhambetov Gabriel J. Brostow

University College London

ICCV 2017

# Target

- Propose a simple method to construct a deep feature space, with explicitly disentangled representations of several known transformations.

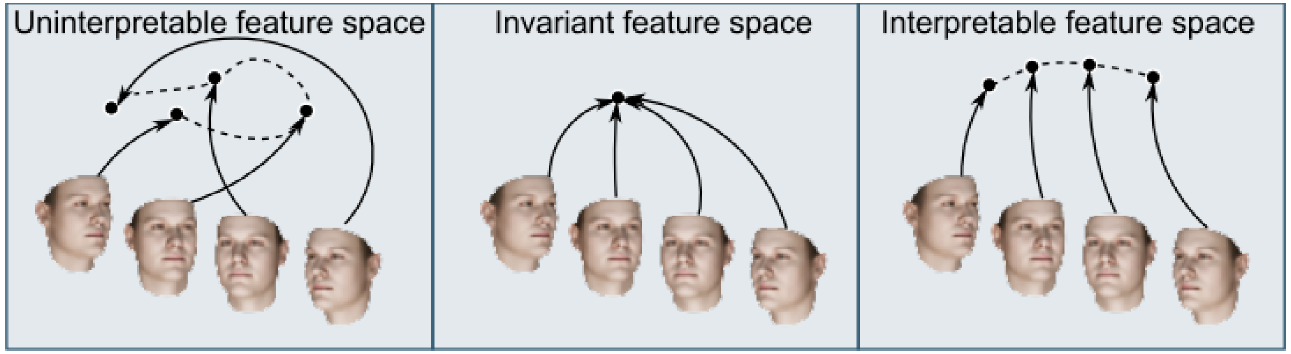



Figure 1. Three alternative feature spaces and how each encodes images of the same person. (Left) A feature space that is hard to interpret, similar to one learned by a typical CNN. While transformation information is present, it is not obvious how to extract that directly from the feature space. (Middle) A transformation-*invariant* feature space. (Right) An interpretable feature-space, where ordered transformations of the input subject relate to ordered, structured features. This is like a learned metric space, but also allows for image synthesis. Images of another person are not shown, but would ideally project similarly, albeit elsewhere in each feature space.

# Method

- Problem setup:

$$\mathcal{D} = \{(\mathbf{x}^1, \tilde{\mathbf{x}}^1_{\theta^i}, \theta^1), ..., (\mathbf{x}^N, \tilde{\mathbf{x}}^N_{\theta^i}, \theta^N)\}$$

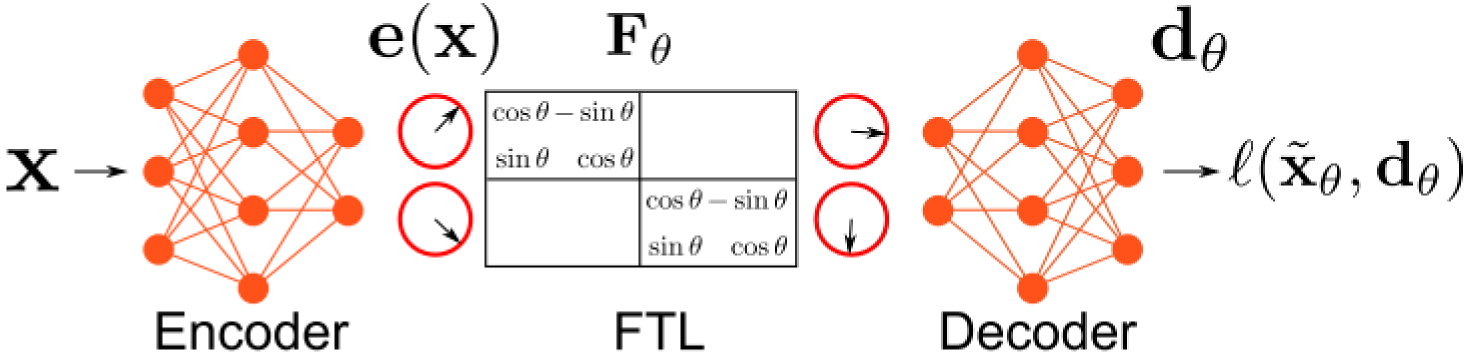
$$\tilde{\mathbf{x}}_\theta = \Pi[\mathcal{T}_\theta[\mathbf{o}]] = \Pi\left[\mathcal{T}_\theta\left[\Pi^{-1}[\mathbf{x}]\right]\right]$$

$$\tilde{\mathbf{x}}_\theta = \mathbf{d}\left(\mathcal{F}_{\theta^i}\left[\mathbf{e}(\mathbf{x}^i)\right]\right)$$

$\mathbf{e}(\bullet)$ approximates $\Pi^{-1}[\bullet]$,

$\mathcal{F}_\theta$ is the feature space equivalent to $\mathcal{T}_\theta$,

$\mathbf{d}(\bullet)$ approximates $\Pi[\bullet]$,



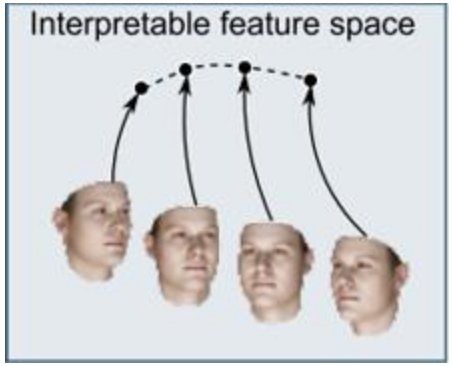
# The Feature Transform Layer

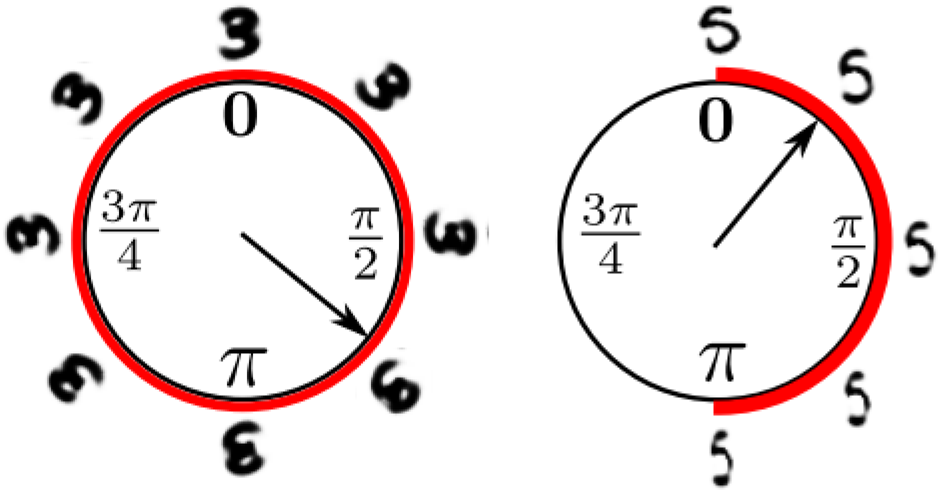
$$\mathbf{y} = \mathcal{F}_\theta[\mathbf{e}] = \mathbf{F}_\theta \mathbf{e}.$$

$$\mathbf{F}_{\theta_2\theta_1} = \mathbf{F}_{\theta_2}\mathbf{F}_{\theta_1}.$$

$$\mathbf{F}_{\theta_1^{-1}} = \mathbf{F}_{\theta_1}^{-1}.$$

$$\|\mathbf{R}_\theta \mathbf{e}\|_2^2 = \mathbf{e}^\top \mathbf{R}_\theta^\top \mathbf{R}_\theta \mathbf{e} = \mathbf{e}^\top \mathbf{e} = \|\mathbf{e}\|_2^2,$$

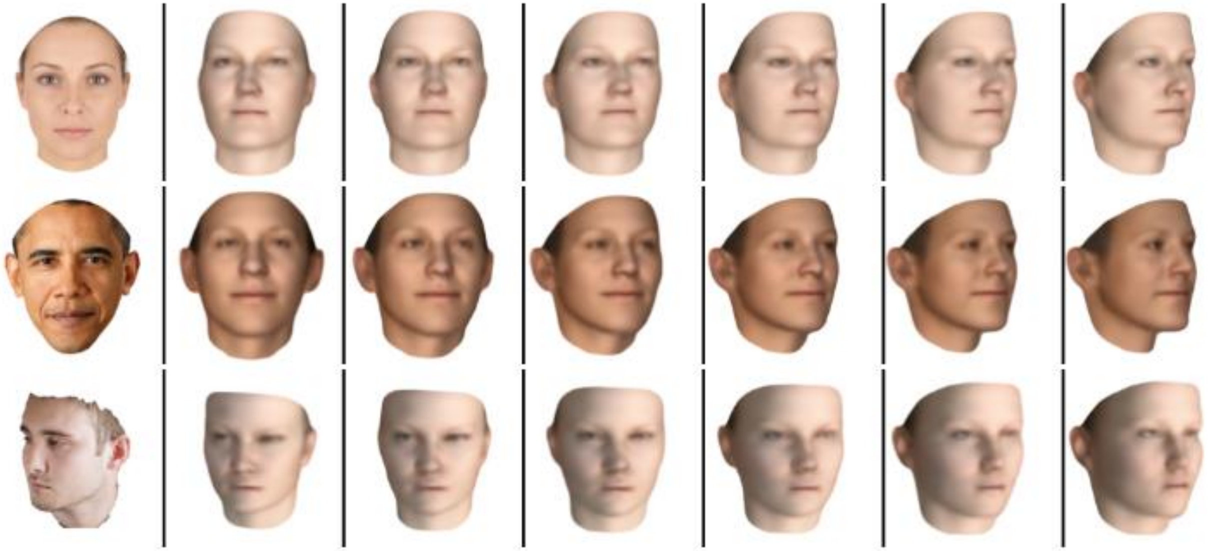


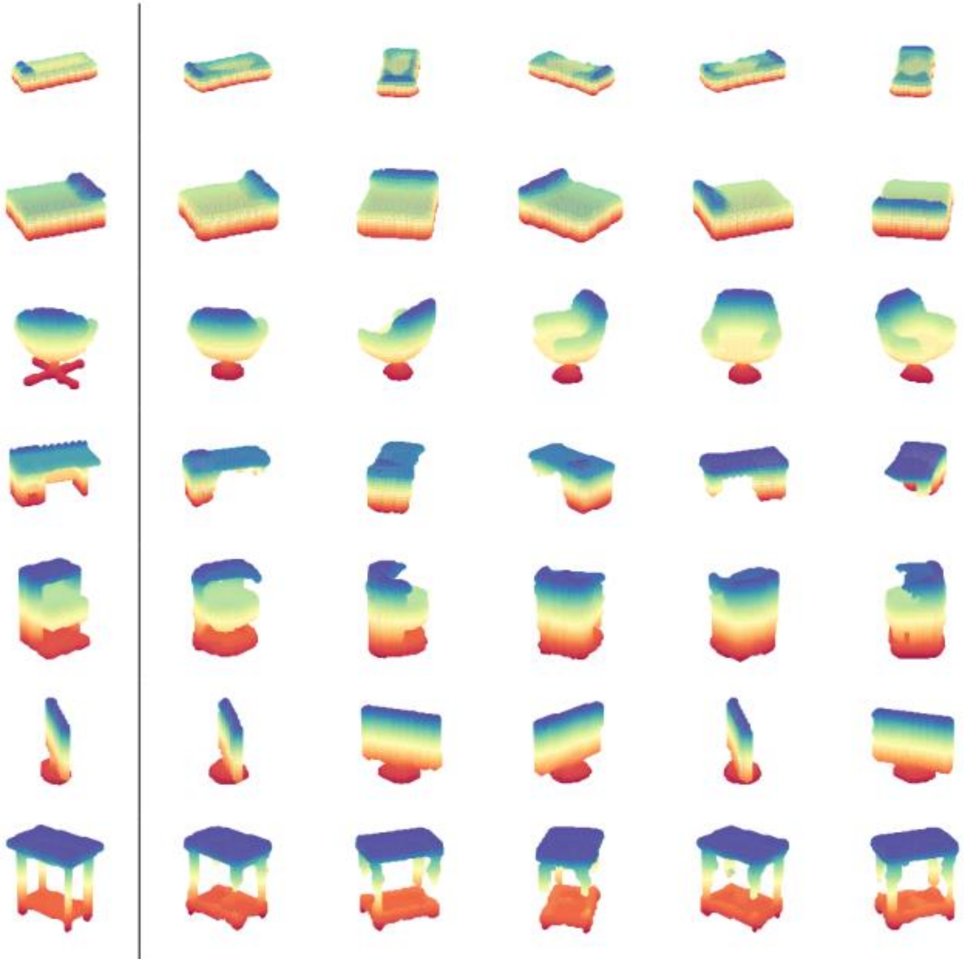




$$\mathbf{F}_\theta \mathbf{e} = \begin{bmatrix} \mathbf{R}_{\theta_1} & & \\ & \ddots & \\ & & \mathbf{R}_{\theta_N} \end{bmatrix} \mathbf{e},$$

对应变换的不同自由度

# Result



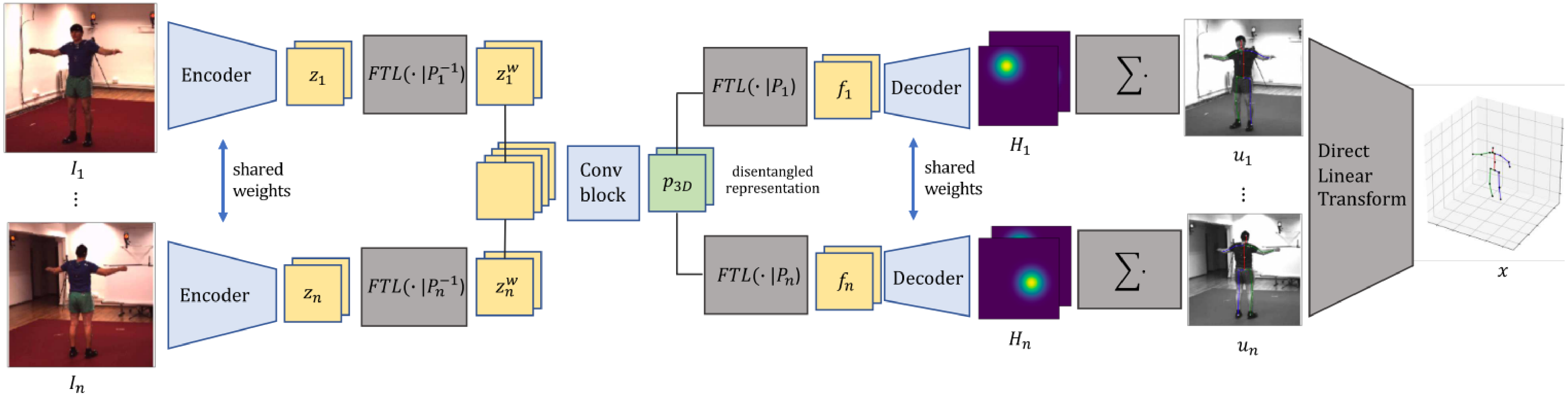


# Lightweight Multi-View 3D Pose Estimation through Camera-Disentangled Representation

**Edoardo Remelli; Shangchen Han; Sina Honari; Pascal Fua; Robert Wang**

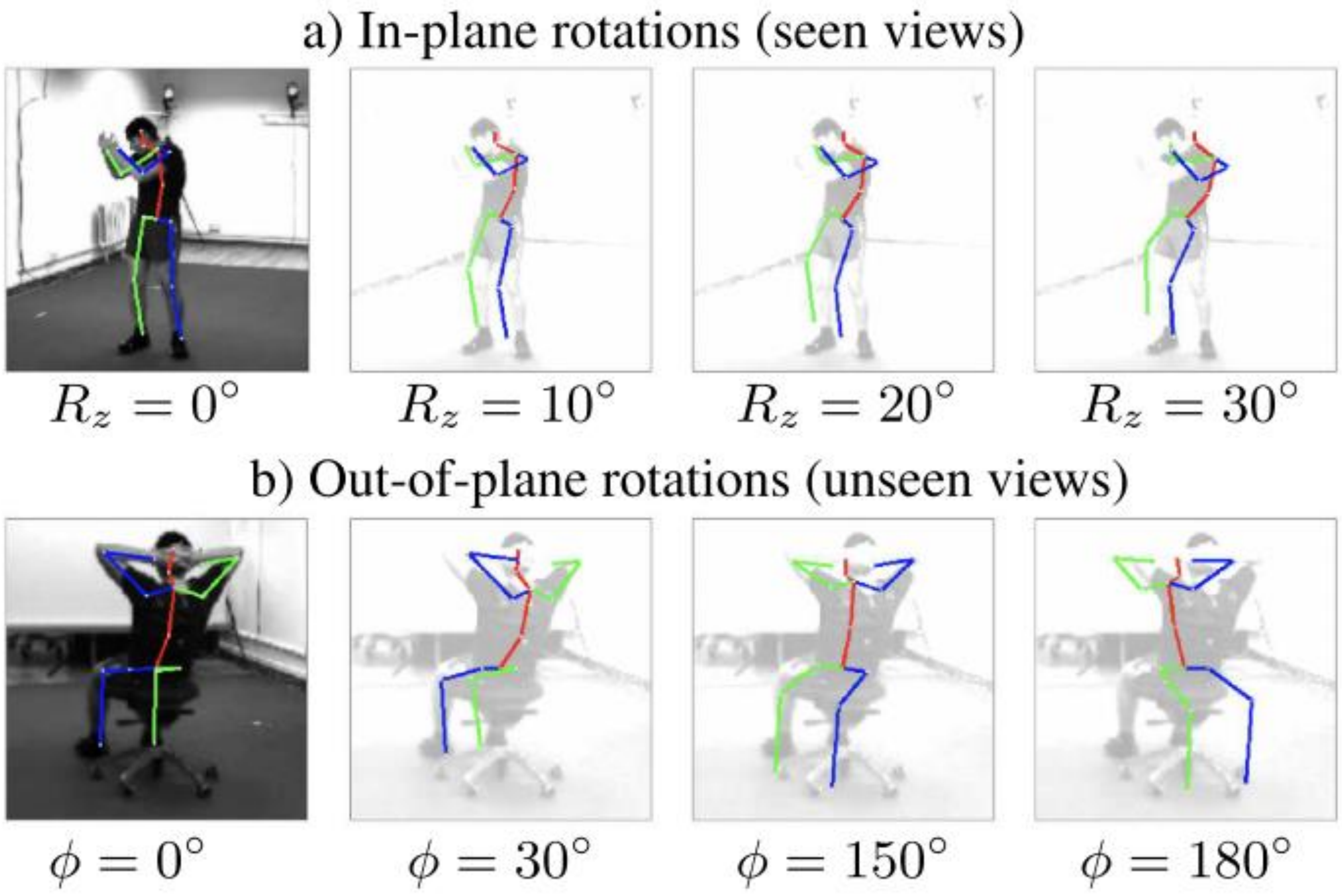
**CVLab, EPFL, Lausanne, Switzerland; Facebook Reality Labs, Redmond, USA**

CVPR 2020



Encoder
$z_1$
$FTL(\cdot|P_1^{-1})$
$z_1^w$
$I_1$
shared weights
Encoder
$z_n$
$FTL(\cdot|P_n^{-1})$
$z_n^w$
$I_n$
Conv block
$p_{3D}$
disentangled representation
$FTL(\cdot|P_1)$
$f_1$
Decoder
$H_1$
$\sum\cdot$
$u_1$
shared weights
$FTL(\cdot|P_n)$
$f_n$
Decoder
$H_n$
$\sum\cdot$
$u_n$
Direct Linear Transform
$x$

# Result


a) In-plane rotations (seen views)
$R_z = 0°$
$R_z = 10°$
$R_z = 20°$
$R_z = 30°$
b) Out-of-plane rotations (unseen views)
$\phi = 0°$
$\phi = 30°$
$\phi = 150°$
$\phi = 180°$