CS 103: Representation Learning, Information Theory and Control

Lecture 2, Jan 18, 2019

Today's program

What is a nuisance for a task?

How do we design nuisance invariant representations?

Invariance, equivariance, canonization

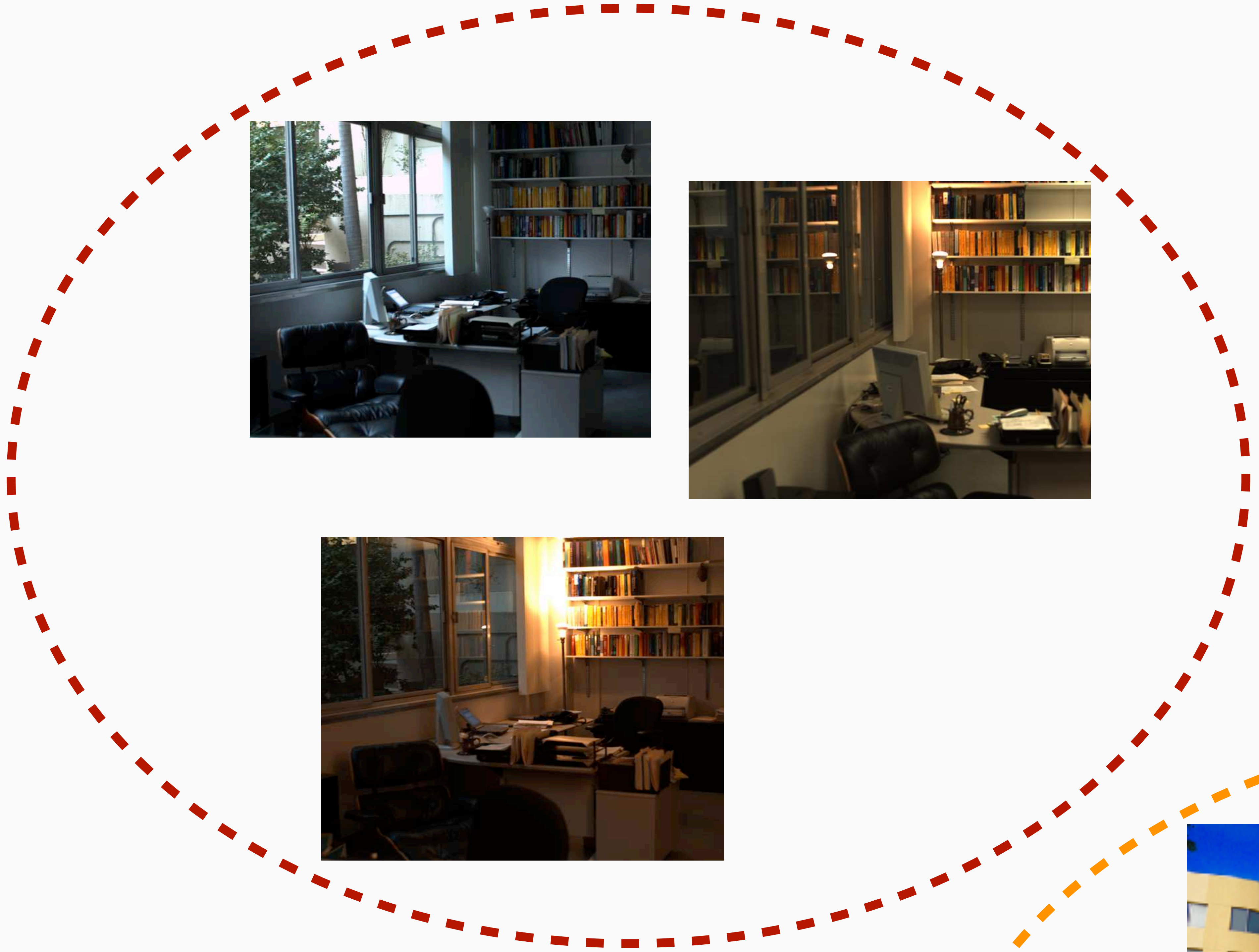
A linear transformation is group equivariant if and only if it is a group convolution

Image canonization with equivariant reference frame detector

Applications to multi-object detection

Nuisance invariance

Why we need nuisance invariance



Images of office from *Steps Toward a Theory of Visual Information*, S. Soatto, 2011

Why we need nuisance invariance

- Office
- Mount Everest
- Team Disneyland Administration

What is a nuisance? It depends on the task

Having different clothes is a nuisance for the task of recognizing the person.

But what if our task is to tag the clothing style in the image?

Pictures of Ian McKellen from https://en.wikipedia.org/wiki/Ian_McKellen

Definition of tasks and nuisances

Let *x* be the input data (*e.g.,* an image), and assume we want to infer the value of a hidden random variable *y* that depends on *x,* that is, we want to reconstruct the posterior distribution $p(y \mid x)$. Then, we call *y* our **task variable**.

Examples:

Image classification: *y* is the label of the image

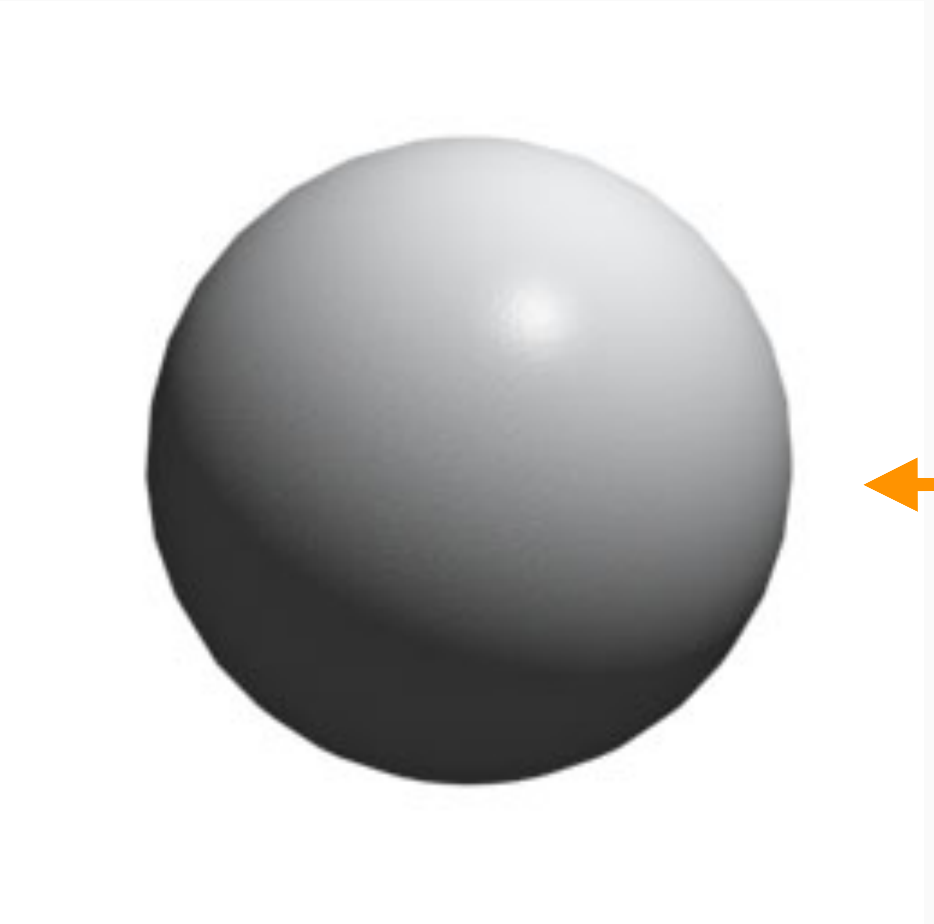
Object detection: *y* is the label and bounding-box of all images in the image

3-D reconstruction: *y* is the 3-D geometry of the scene

Control: *y* is the action to take to bring the system in a certain state

Definition of tasks and nuisances

The observed image *x* may depend on a number of factors. Let's write:



e.g., shape of object

e.g., illumination

$$x = I(\xi, \nu)$$

Rendering function

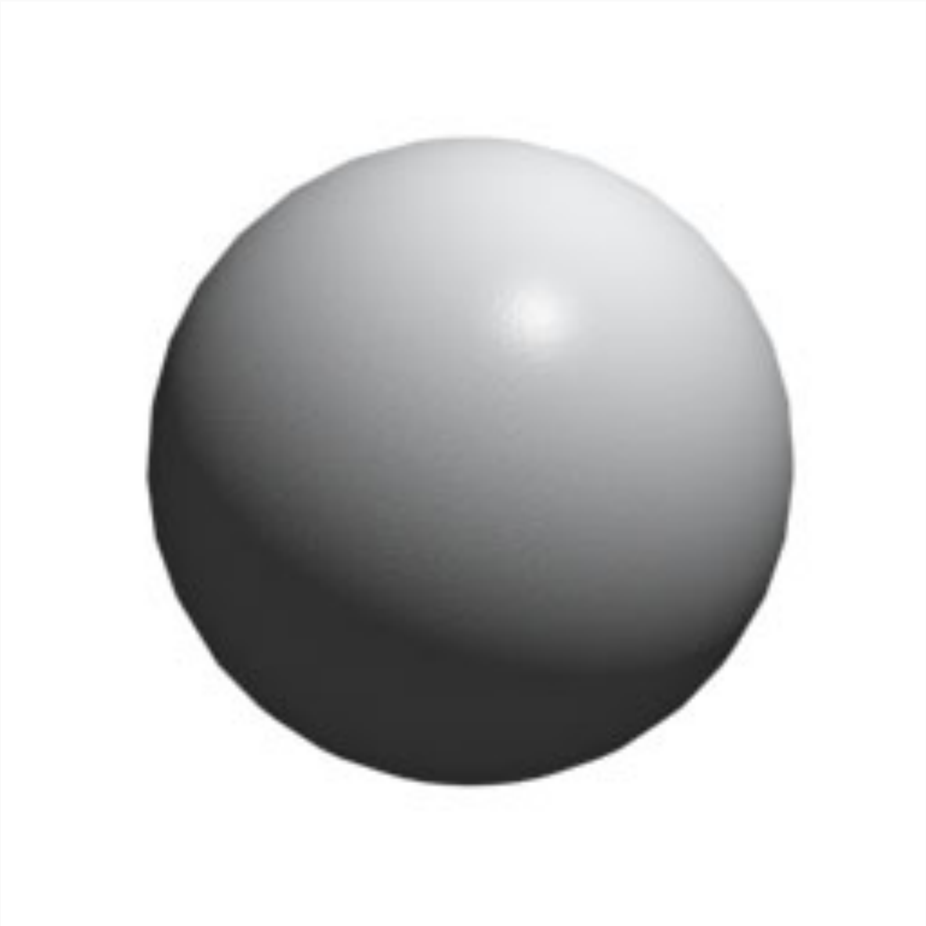
We will prove later that any image distribution can always be parametrized in this way, for an appropriate rendering function *I*.

For now, think of *I* as a powerful and generic photorealistic rendering engine.

Effect of changing the rendering parameters

Effect of changing the parameters of the rendering function.

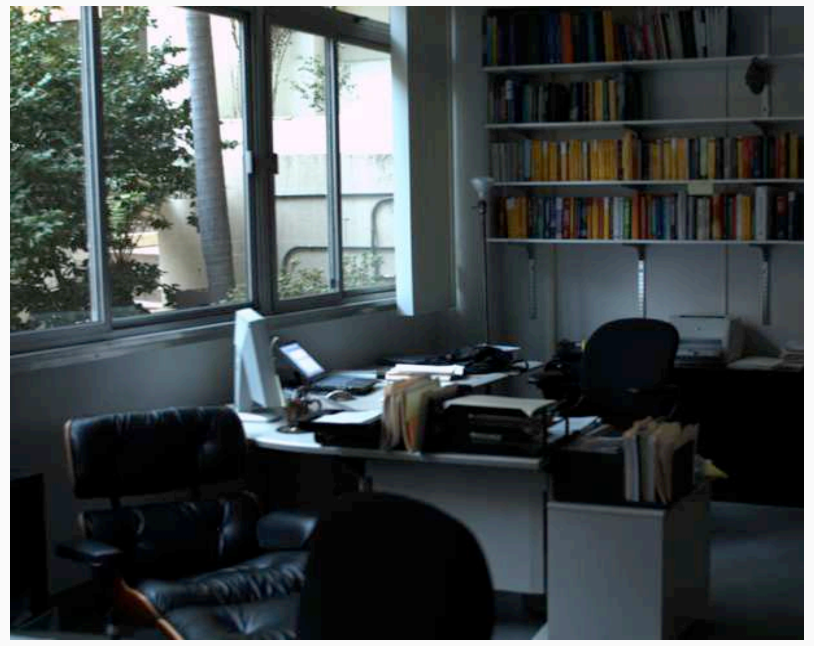
$I(\xi, \nu)$



$I(\xi, \nu')$

$I(\xi', \nu)$

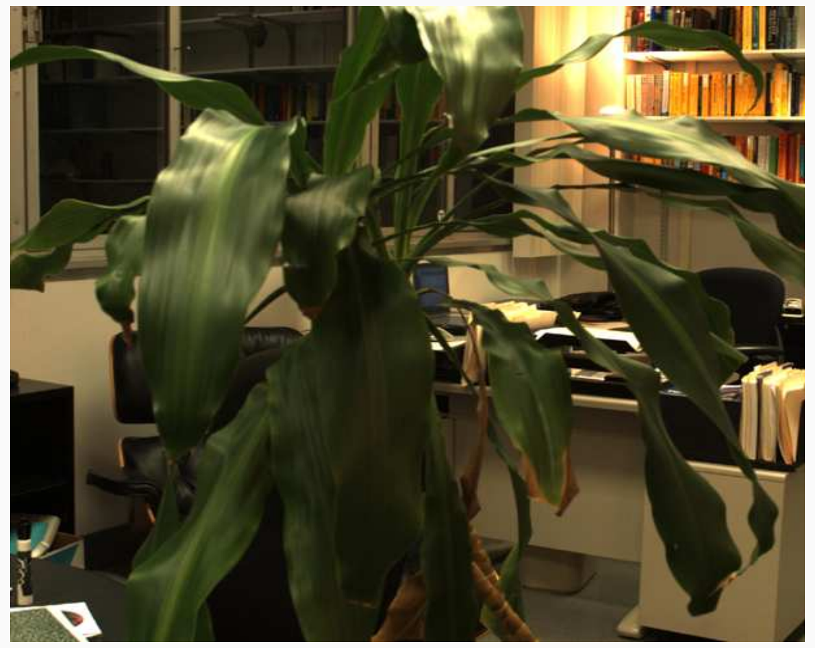
Effect of changing the rendering parameters



$I = h(\xi, \nu)$

Change of illumination, point of view

$\tilde{I} = h(\xi, \tilde{\nu}), \quad \tilde{\nu} = \text{illumination}$



$\tilde{\nu} = \text{visibility}$

$\tilde{\nu} = \text{viewpoint}$

$\tilde{I} = h(\tilde{\xi}, \tilde{\nu}), \quad \tilde{\xi} \neq \xi$

Change of identity

Images from *Steps Toward a Theory of Visual Information*, S. Soatto, 2011

Definition of nuisance

Suppose that changing $\boldsymbol{\nu}$ does not affect the task variable *y*. That is:

$$p(y \,|\, I(\xi, \nu)) = p(y \,|\, I(\xi, \nu')) \textbf{ for all } \nu' \in N$$

Then we say that $\boldsymbol{\nu}$ is a **nuisance** *for the task y.*

Note: This is equivalent to saying that y is independent of $\boldsymbol{\nu}$, or alternatively that $\boldsymbol{\nu}$ contains no information about the task *y*, *i.e.*, $I(y; \boldsymbol{\nu}) = 0$

Common examples:

Illumination, change of contrast, rotations, translations, change of scale, …

Nuisance invariance

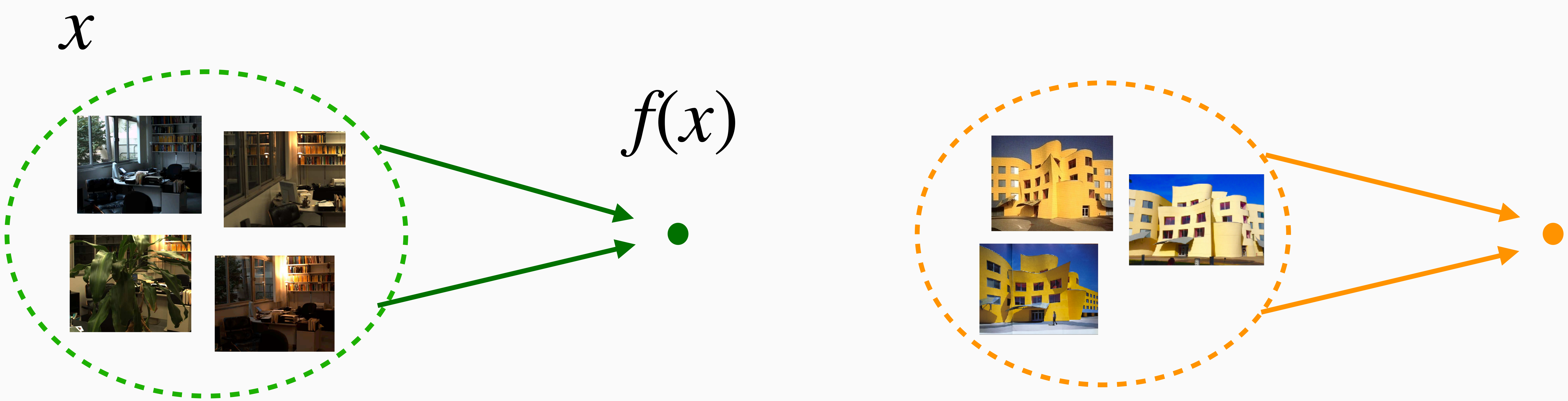
We say that a representation *z* = *f(x)* is **nuisance invariant** if:

$$f(I(\xi,\nu)) = f(I(\xi,\nu'))$$

For all nuisances $\boldsymbol{\nu}$ and $\boldsymbol{\nu}$'.

A representation is *maximal invariant* if all other invariant representations are a function of it.

Idea: a nuisance invariant representation z throws away unneeded information.



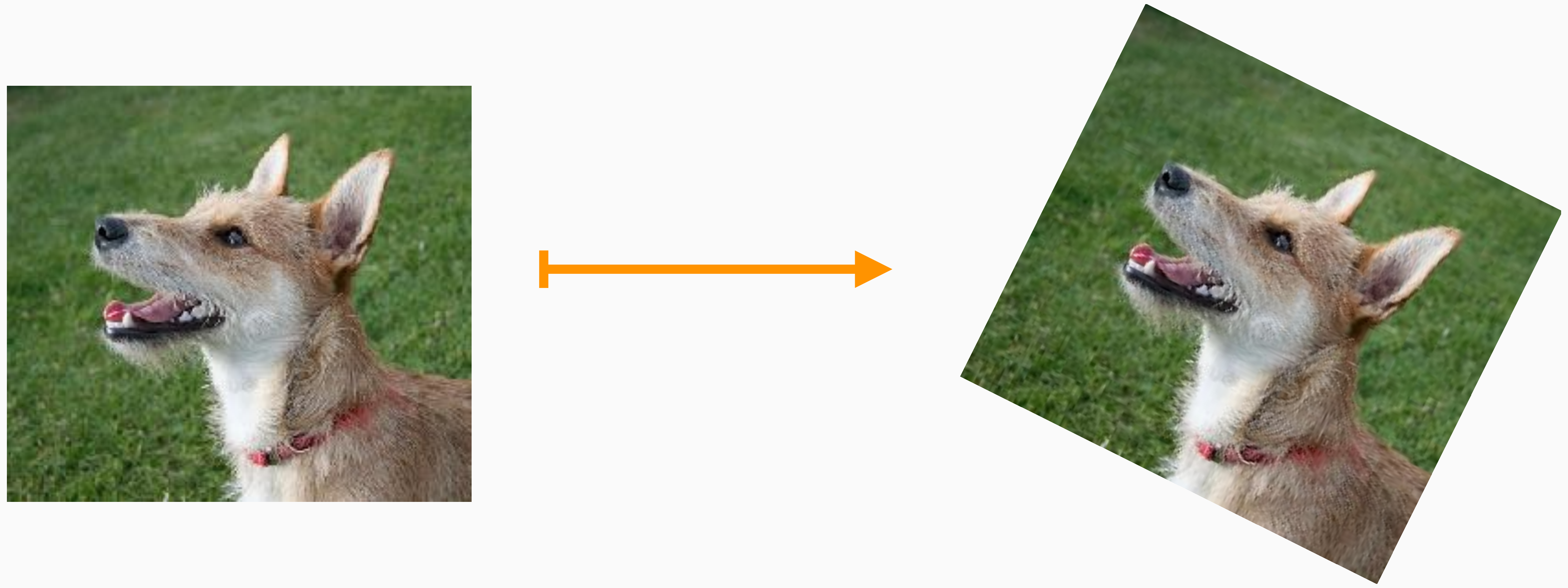
How do we design (maximal) invariant representations?

Far from trivial in the general case.

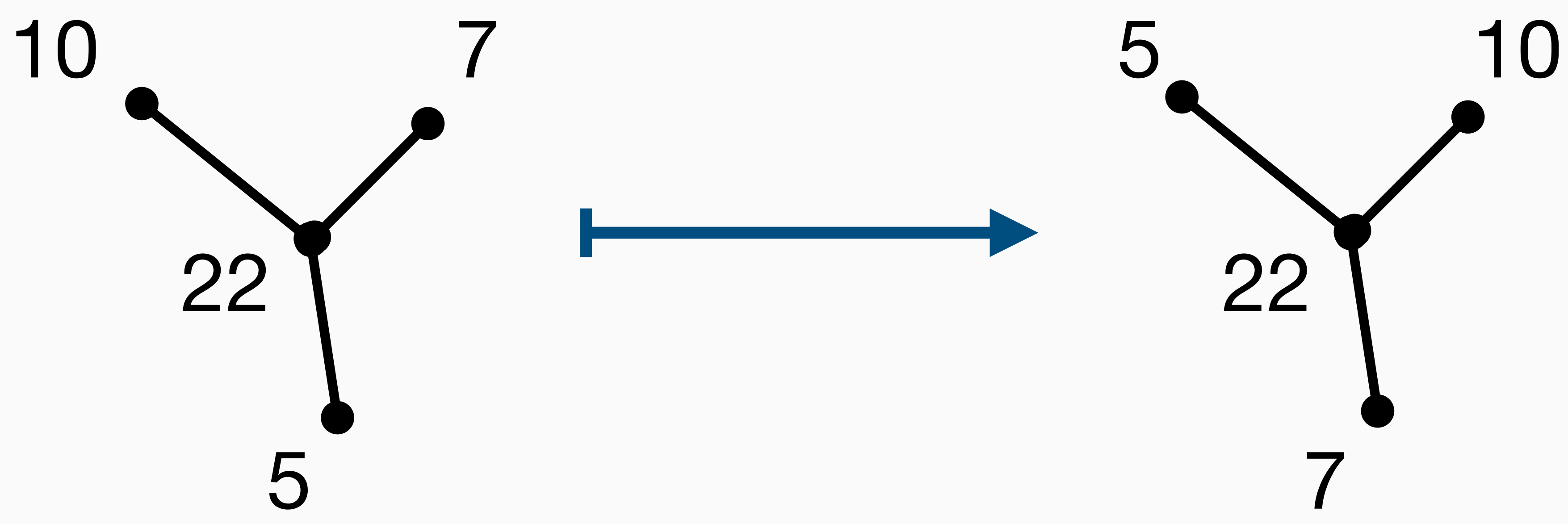
For simple (but important!) group nuisances we can develop a theory.

$$I(\xi, \nu') = g_{\nu \to \nu'} \circ I(\xi, \nu)$$

Translations, rotations



Permutation of vertexes



Group nuisances

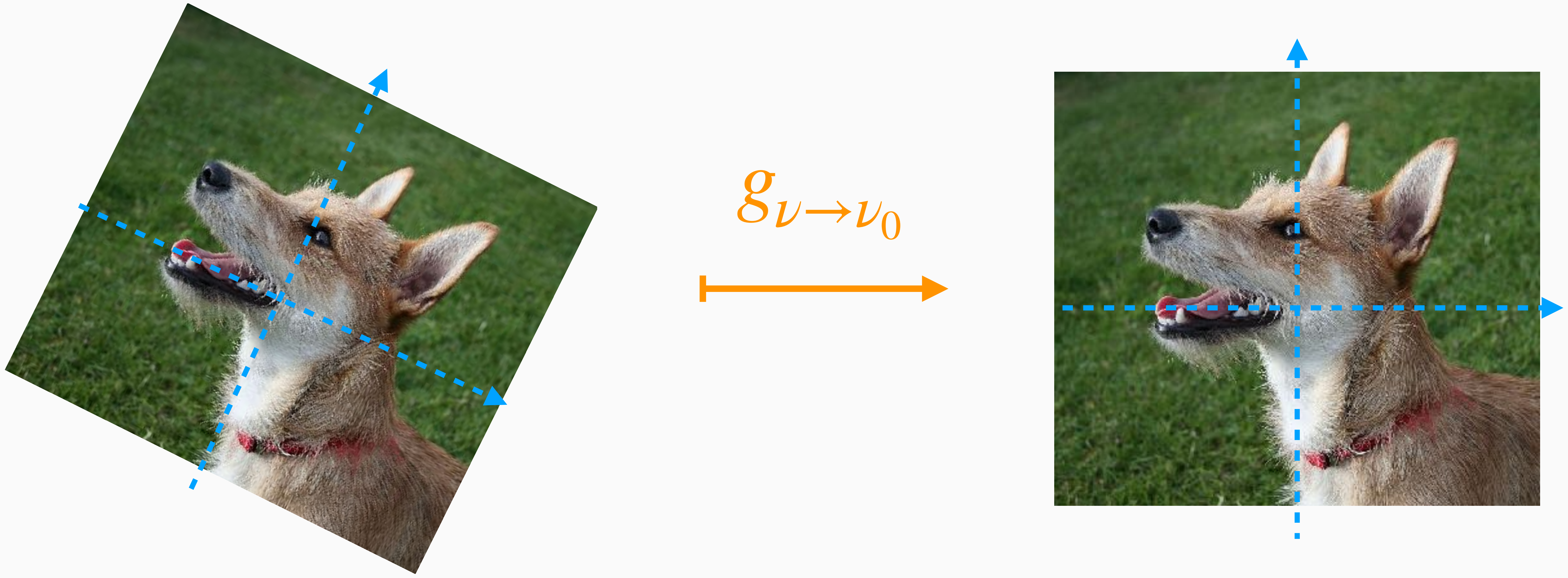
This part was done on whiteboard, see LaTeX notes on class website.

Canonization

Invariance by canonization

Idea: Instead of finding an invariant representation, apply a transformation to put the input in a standard form.

$$I(\xi,\nu) \longmapsto \; g_{\nu\to\nu_0} \circ I(\xi,\nu) = I(\xi,\nu_0)$$

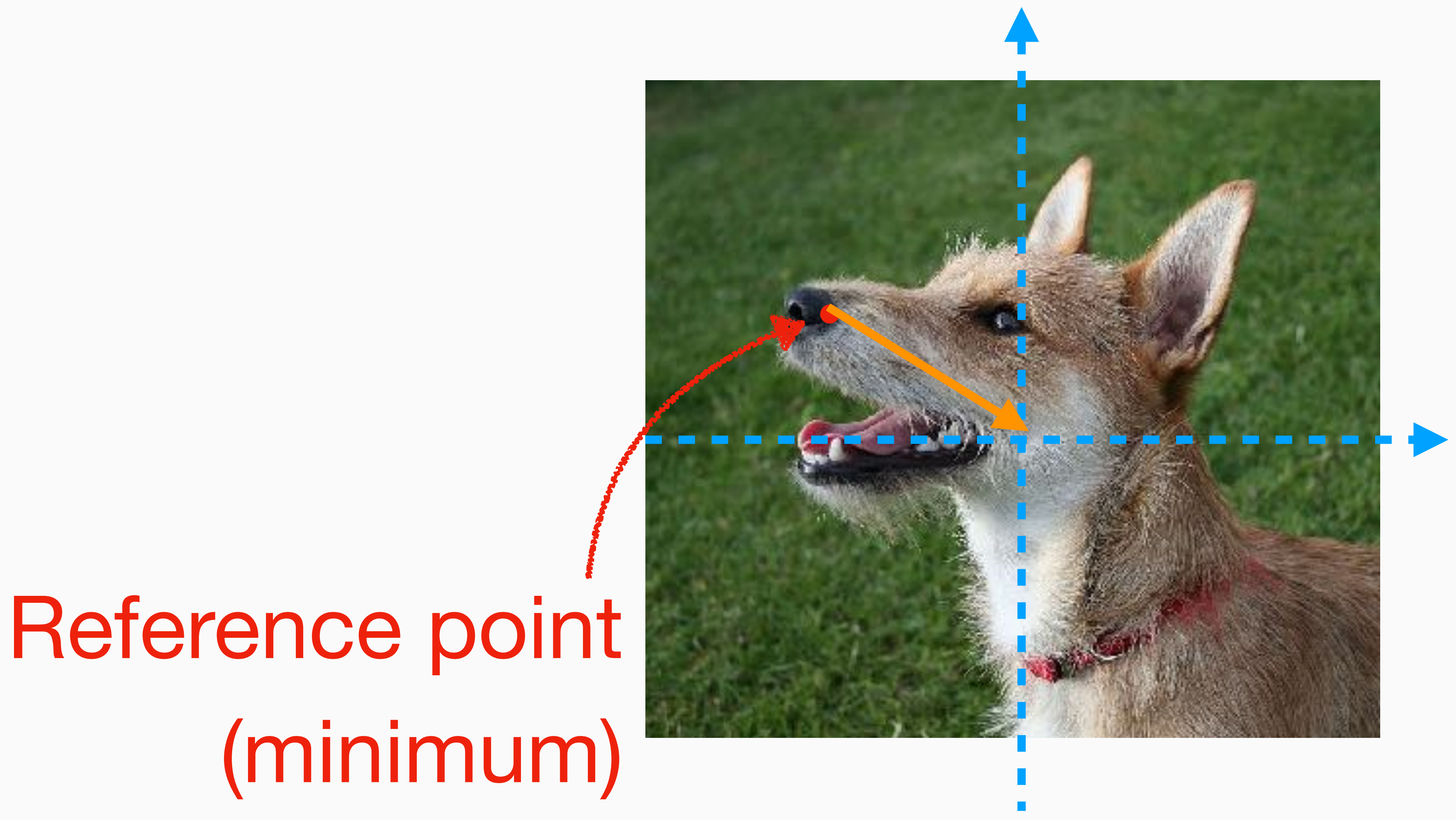


Canonization for translations

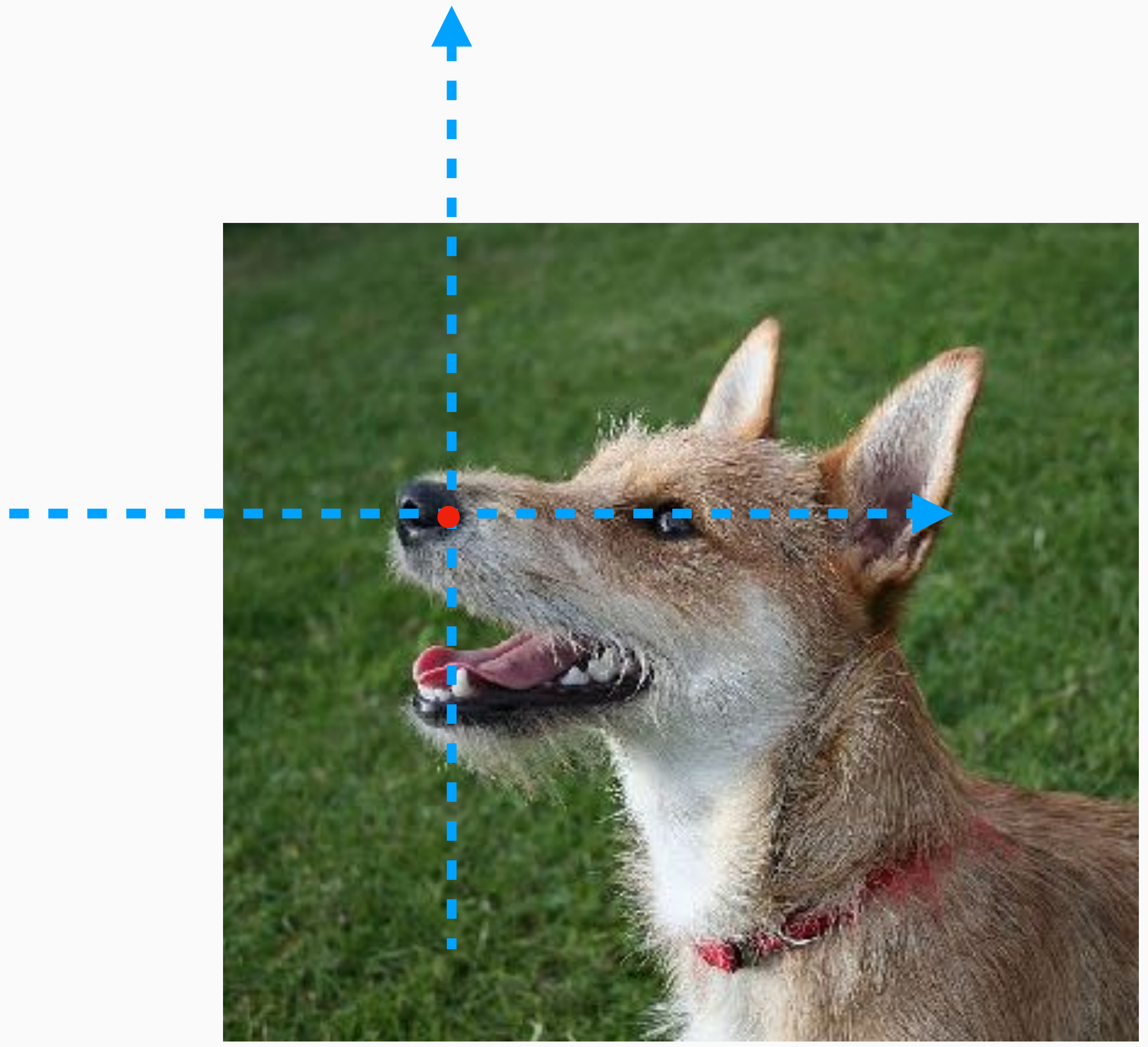
Suppose we want to canonize the image with respect to translations.

1. Decide a reference point that is uniquely defined, no matter how we translate the image

   Examples: The barycenter of the image, the maximum (assuming it's unique)
2. Write an algorithm to find the position of the reference point
3. Compute the translation that moves the reference point to the origin



$g_{\nu' \to \nu_0}$



Equivariant reference frame detector

A reference frame detector R for a group G is any function R(x): $X \to G$ such that

$$R(g \cdot x) = g \cdot R(x)$$

That is, a reference frame detector is any equivariant function from X to G.

Example: Let $G = \mathbf{R}^2$ be the group of translations. Then R(x) = "position of the maximum of x" is a reference frame.

From equivariant frame detector to invariant representations

Proposition. Let *R* be a reference frame detector for the group *G*. Define a representation *f(x) as:*

$$f(x) = R(x)^{-1} \cdot x$$

Then *f(x)* is a *G*-invariant representation.

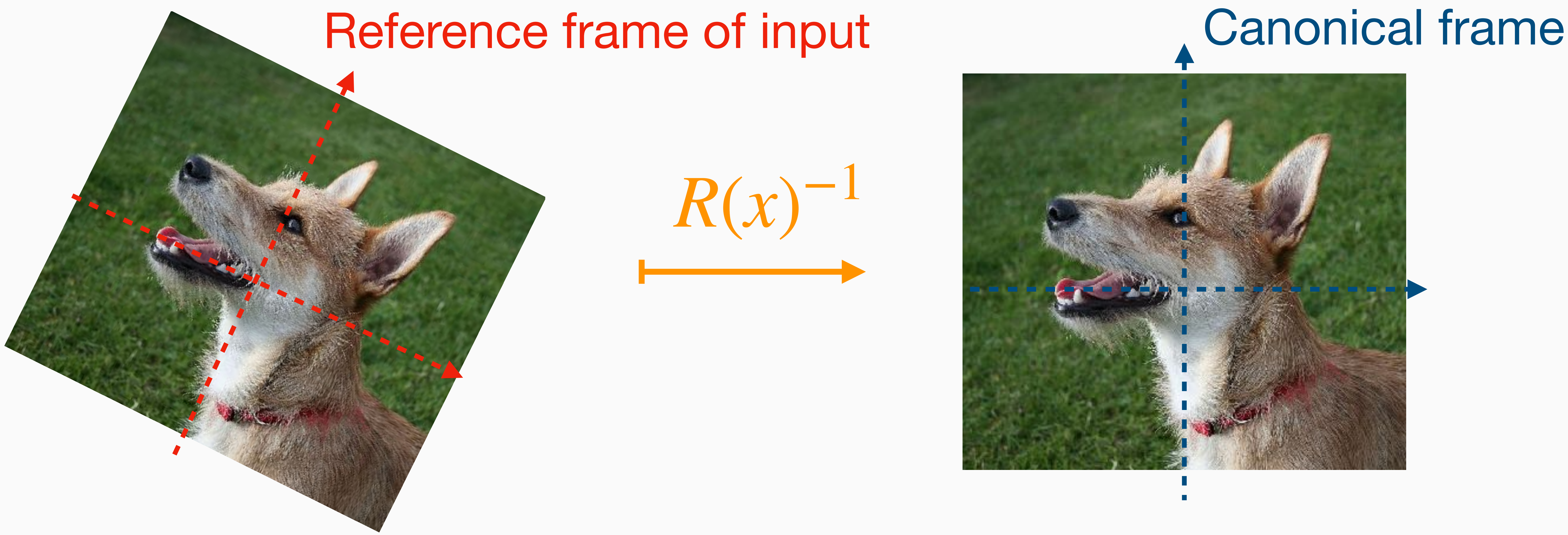
Proof:

$$\begin{aligned} f(g \cdot x) &= R(g \cdot x)^{-1} \cdot (g \cdot x) \\ &= (g \cdot R(x))^{-1} \cdot g \cdot x \\ &= R(x)^{-1} \cdot g^{-1} \cdot g \cdot x \\ &= R(x)^{-1} \cdot x \\ &= f(x) \end{aligned}$$

The canonization pipeline

Canonization consists of the following steps

1. Build an equivariant **reference frame detector**
2. Choose a "**canonical**" reference frame
3. Find the reference frame of the input image
4. Invert the transformation to make the reference frame canonical



Some examples of canonization in vision

Document analysis: Find border of the document and un-warp the image prior to analysis.

Also: Normalize contrast and illumination

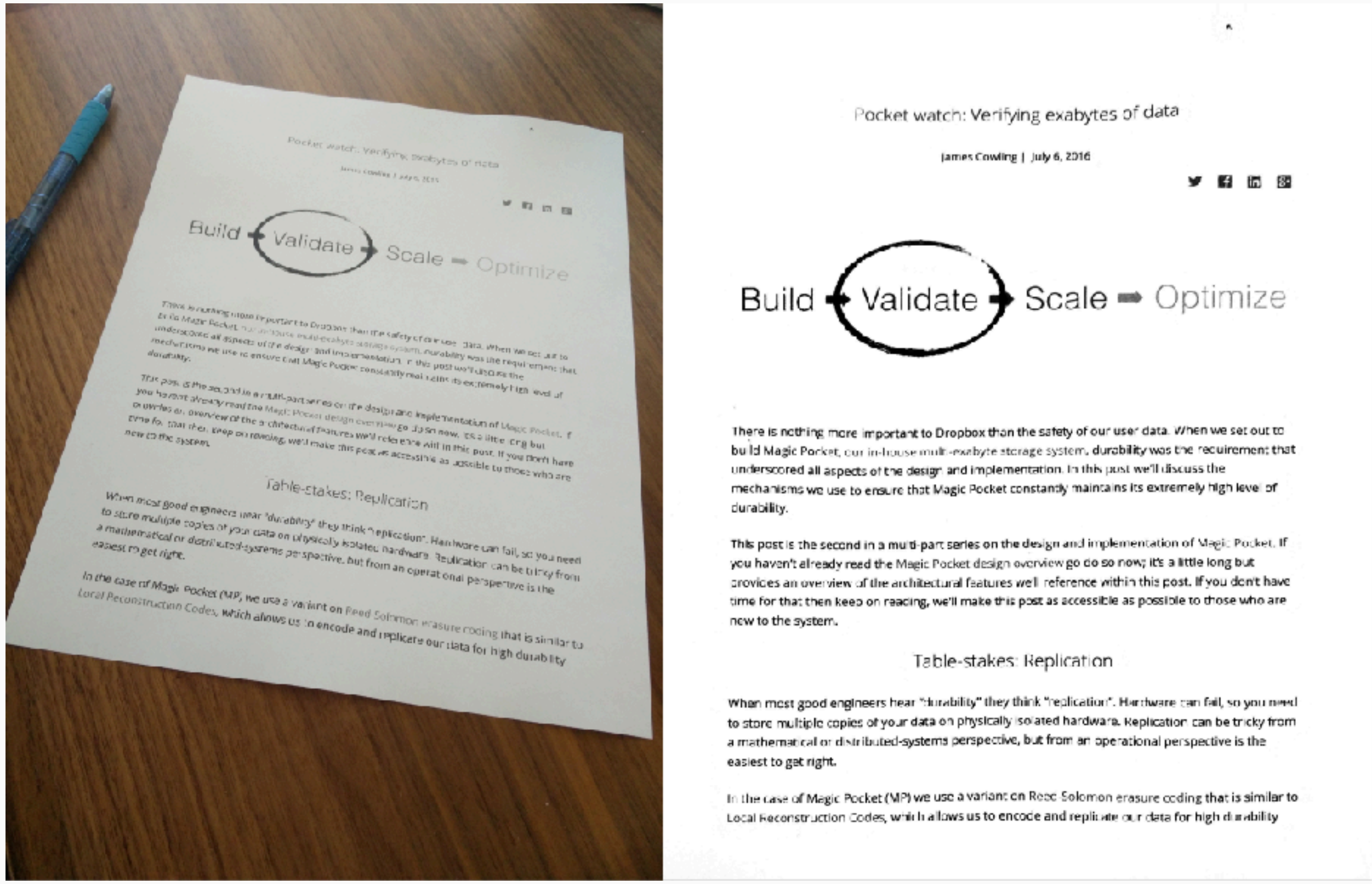
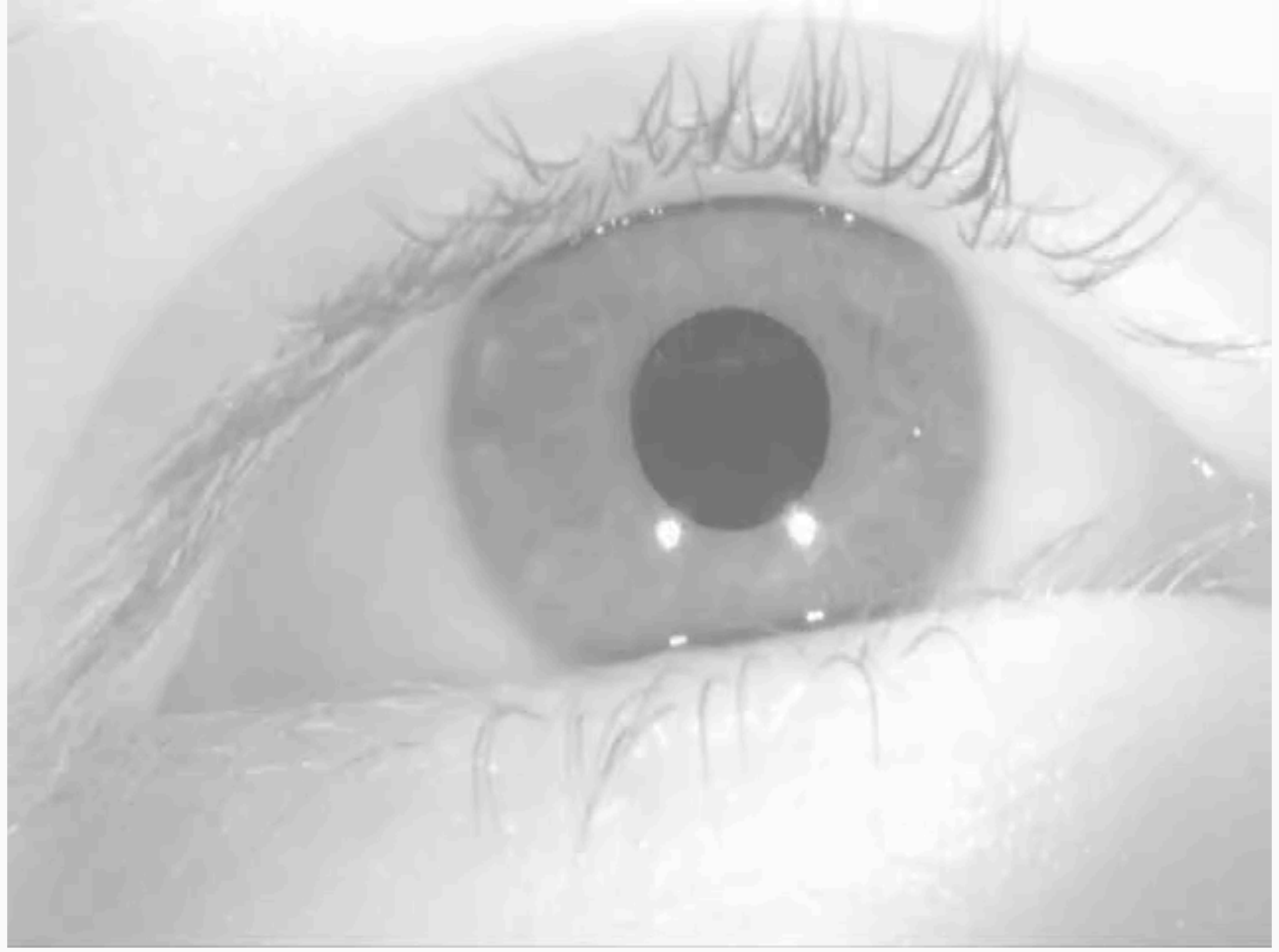


Image from https://blogs.dropbox.com/tech/2016/08/fast-document-rectification-and-enhancement/

Saccades

Eyes move rapidly while looking at a fixed object.



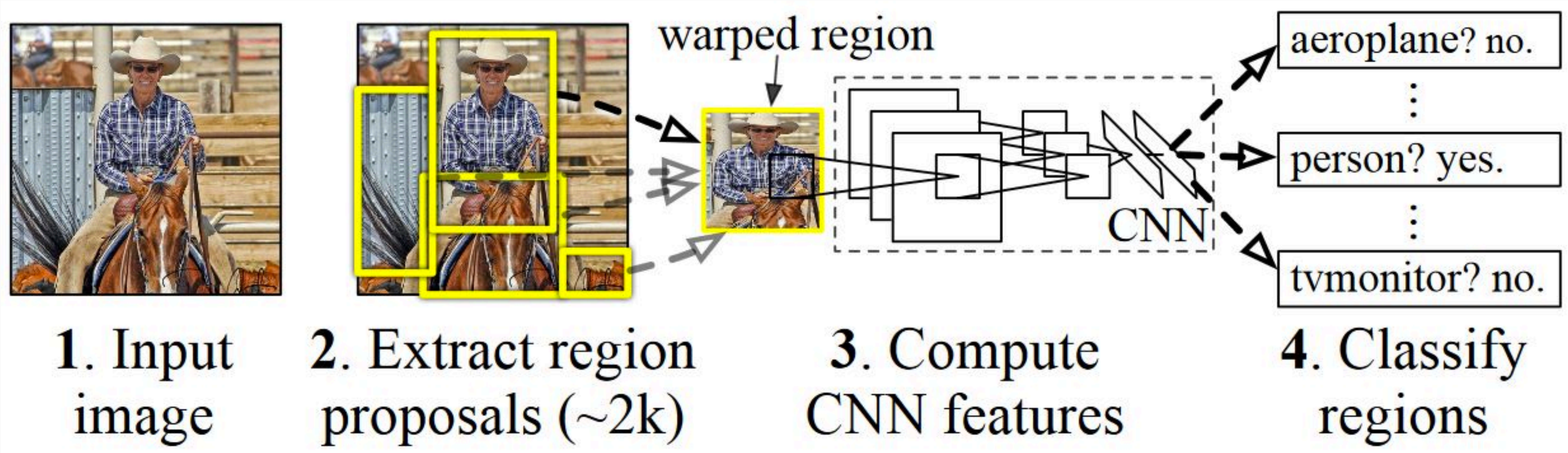
Can we consider this a form of translation invariance by canonization?

Video and Images from https://en.wikipedia.org/wiki/Saccade

The R-CNN model for multi-object detection

Region proposal: find regions of the image that may contain an interesting object (i.e., reference frame proposal)

CNN classifier: warp the region to put it in canonical form (invariance) and feed it to a classifier

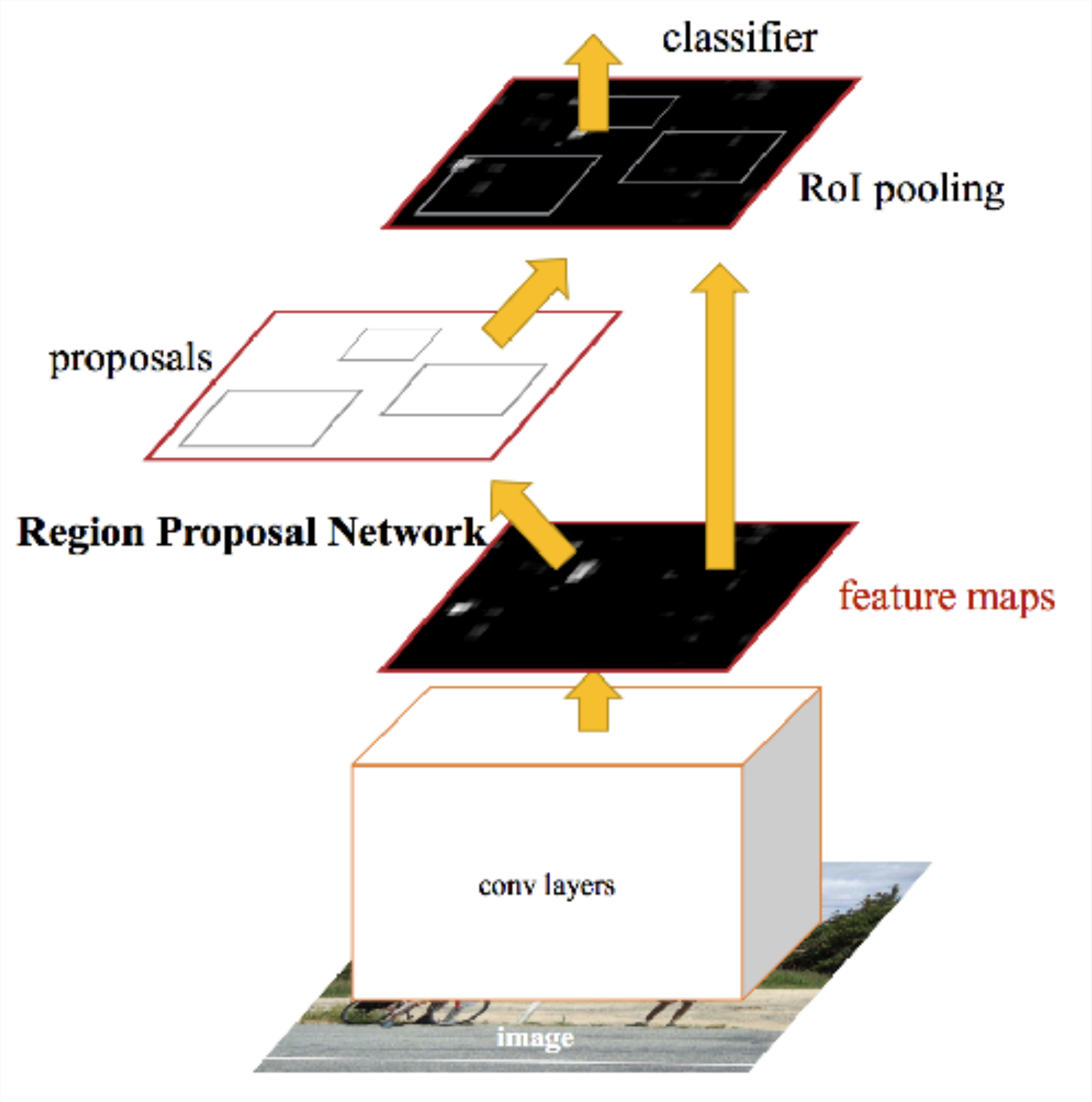


Region proposal + CNN classifier = R-CNN

Region proposal mechanism

Originally: hand-crafted proposal mechanisms based on saliency, uniformity of texture, scale, and so on.

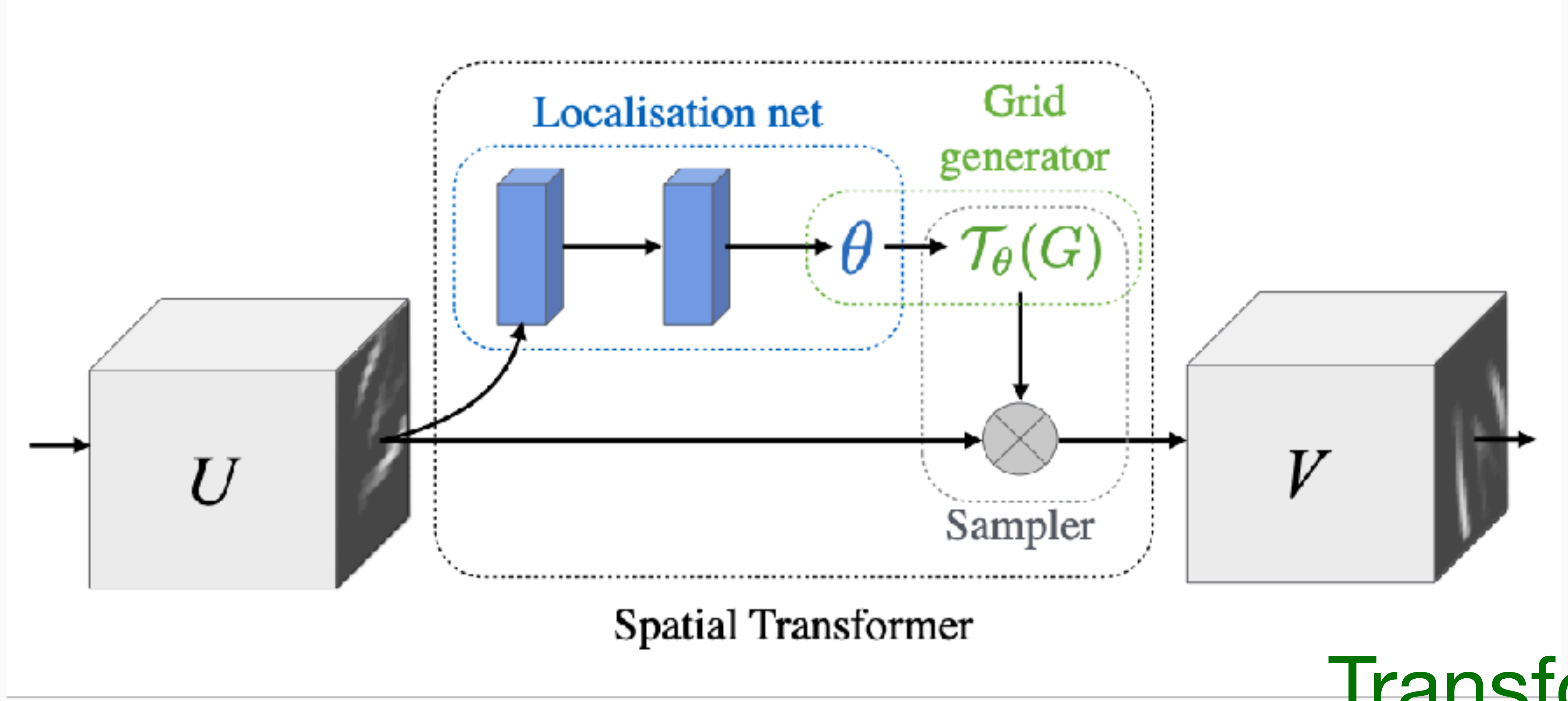
Nowadays: The same network does both the region proposal and the classification inside each region

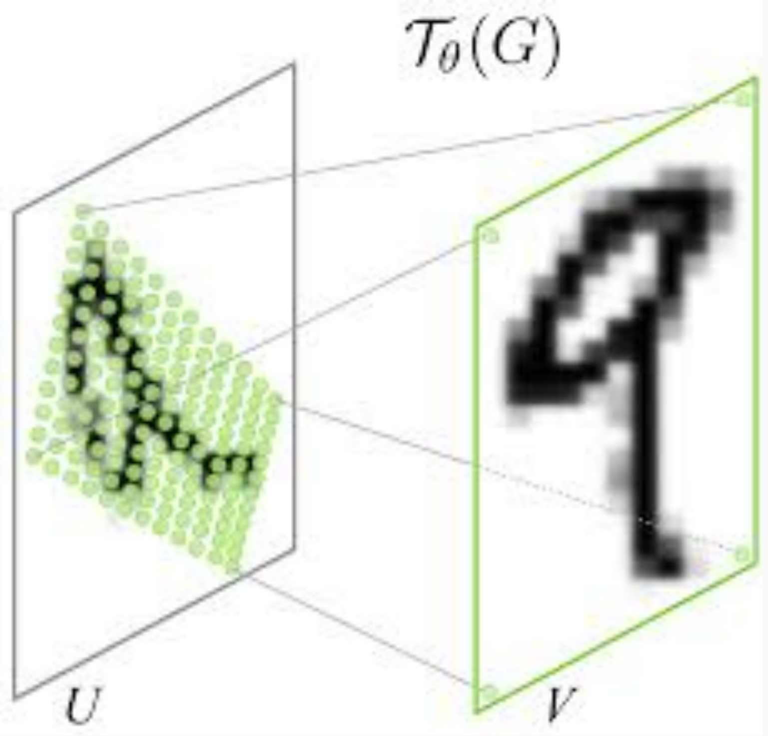


Fast R-CNN

Spatial Transformer Network

Localisation network selects a local reference frame in the image





Transformer resamples using that reference frame