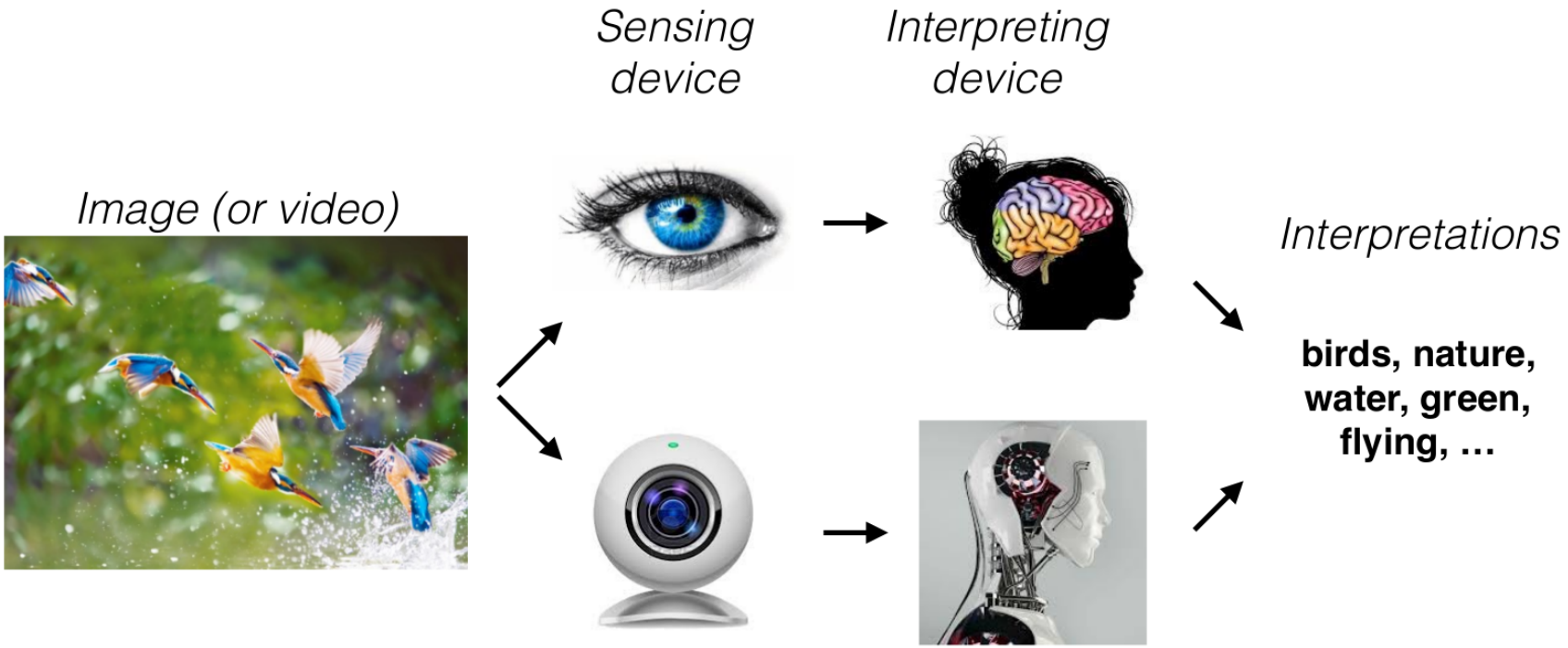
Australian National University

Visual Recognition From Structured Supervision

Rodrigo Santa Cruz and Stephen Gould

Visual Recognition



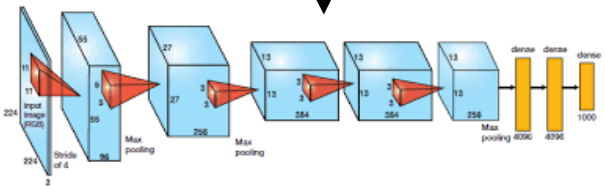
Visual Recognition in the Deep Learning Era

Training Data

person

car

Test Data



Person
Car

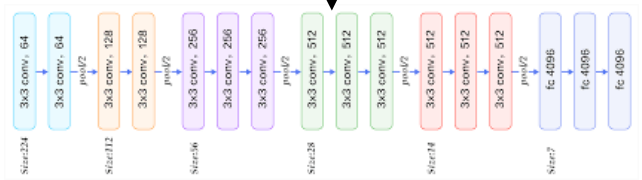
Visual Recognition in the Deep Learning Era

Training Data

person

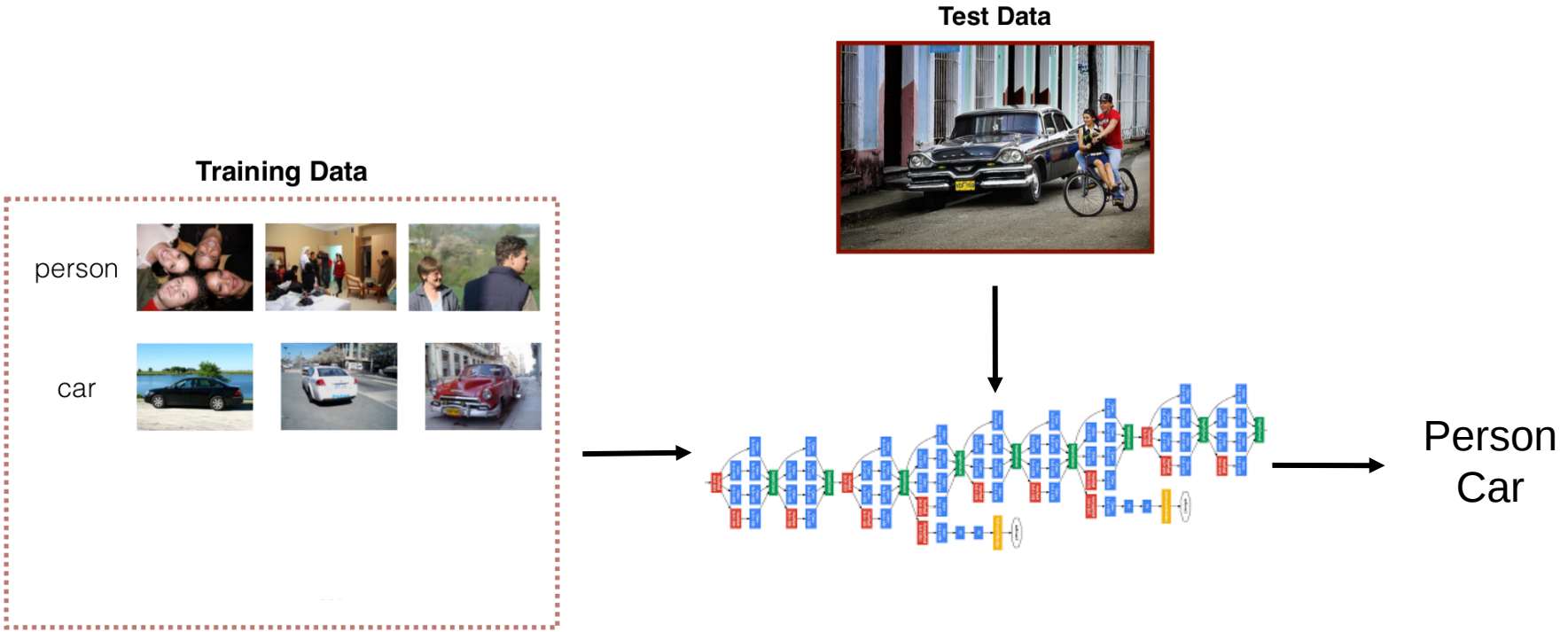
car

Test Data

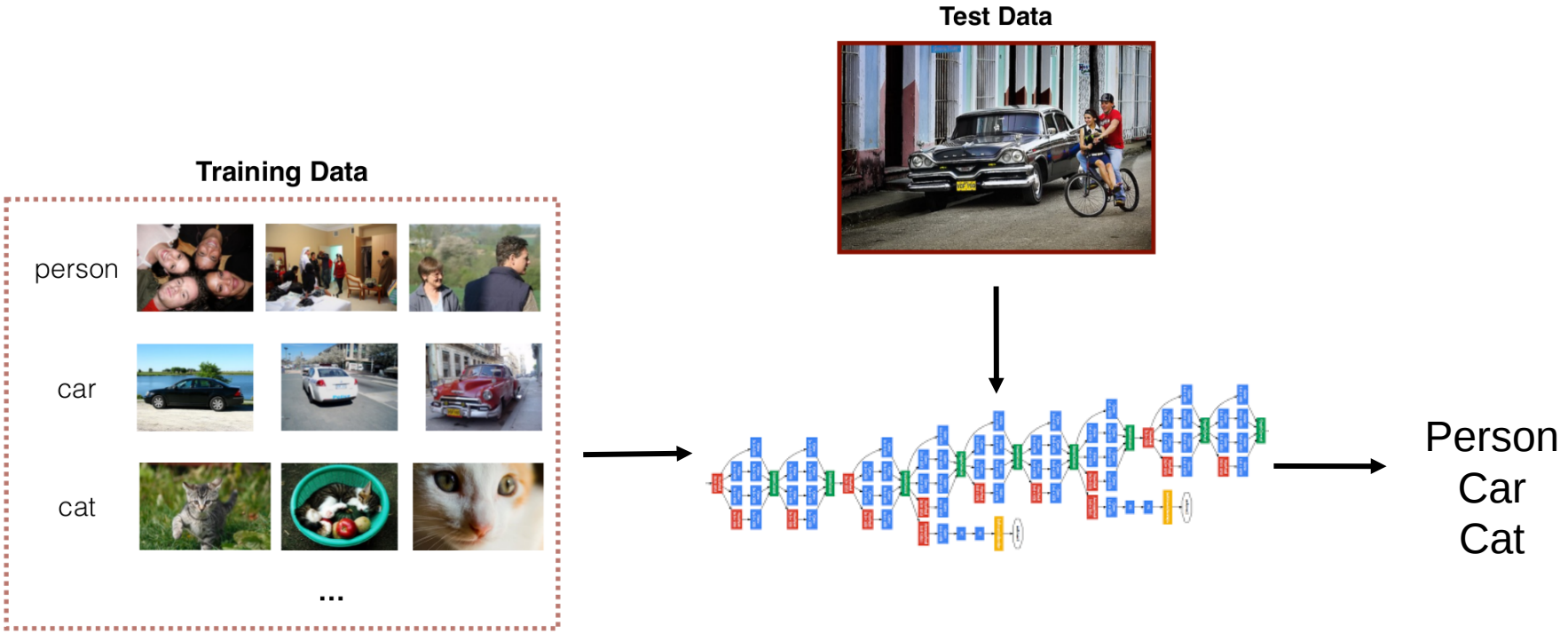


Person
Car

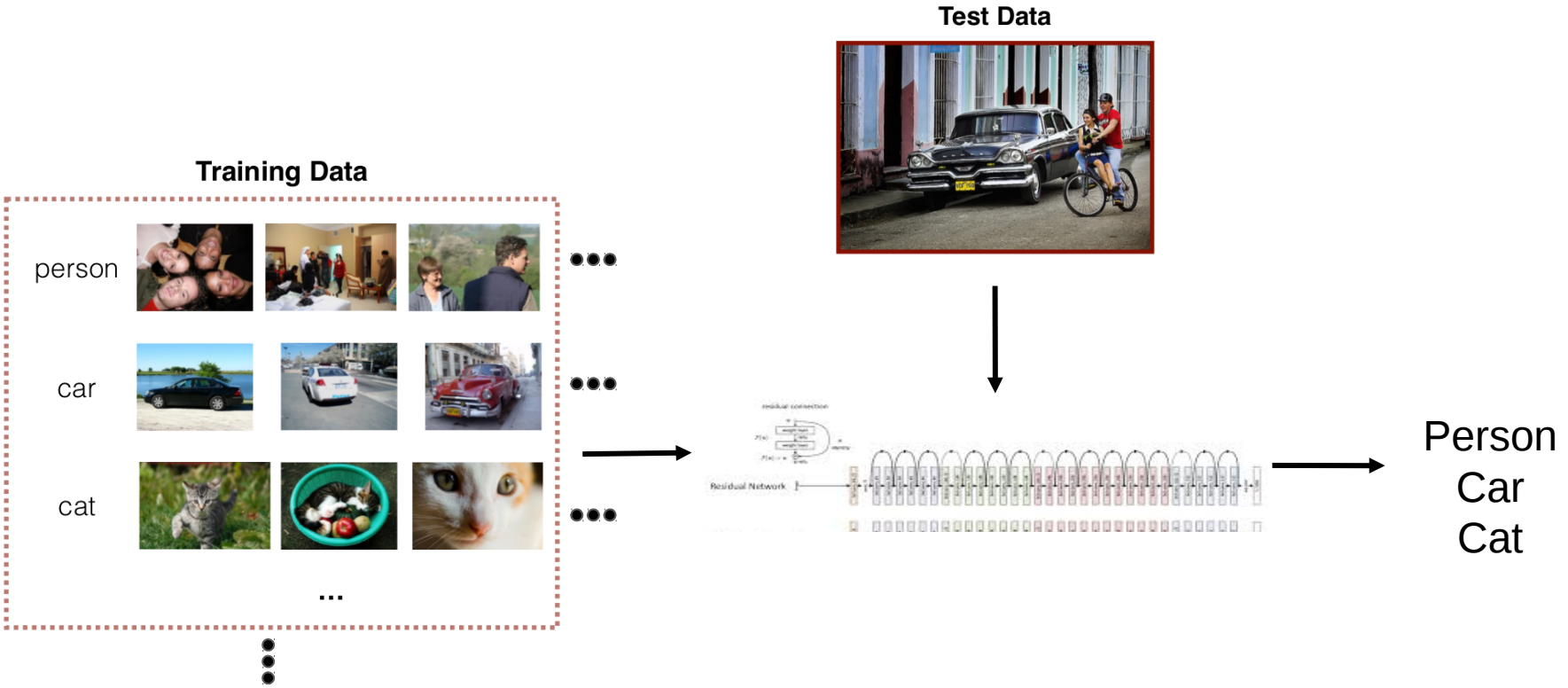
Visual Recognition in the Deep Learning Era



Visual Recognition in the Deep Learning Era



Visual Recognition in the Deep Learning Era



152 layers

22 layers

19 layers

8 layers

8 layers

shallow

Human error (5.1)

3.57

6.7

7.3

11.7

16.4

25.8

28.2

ILSVRC'15 ResNet

ILSVRC'14 GoogleNet

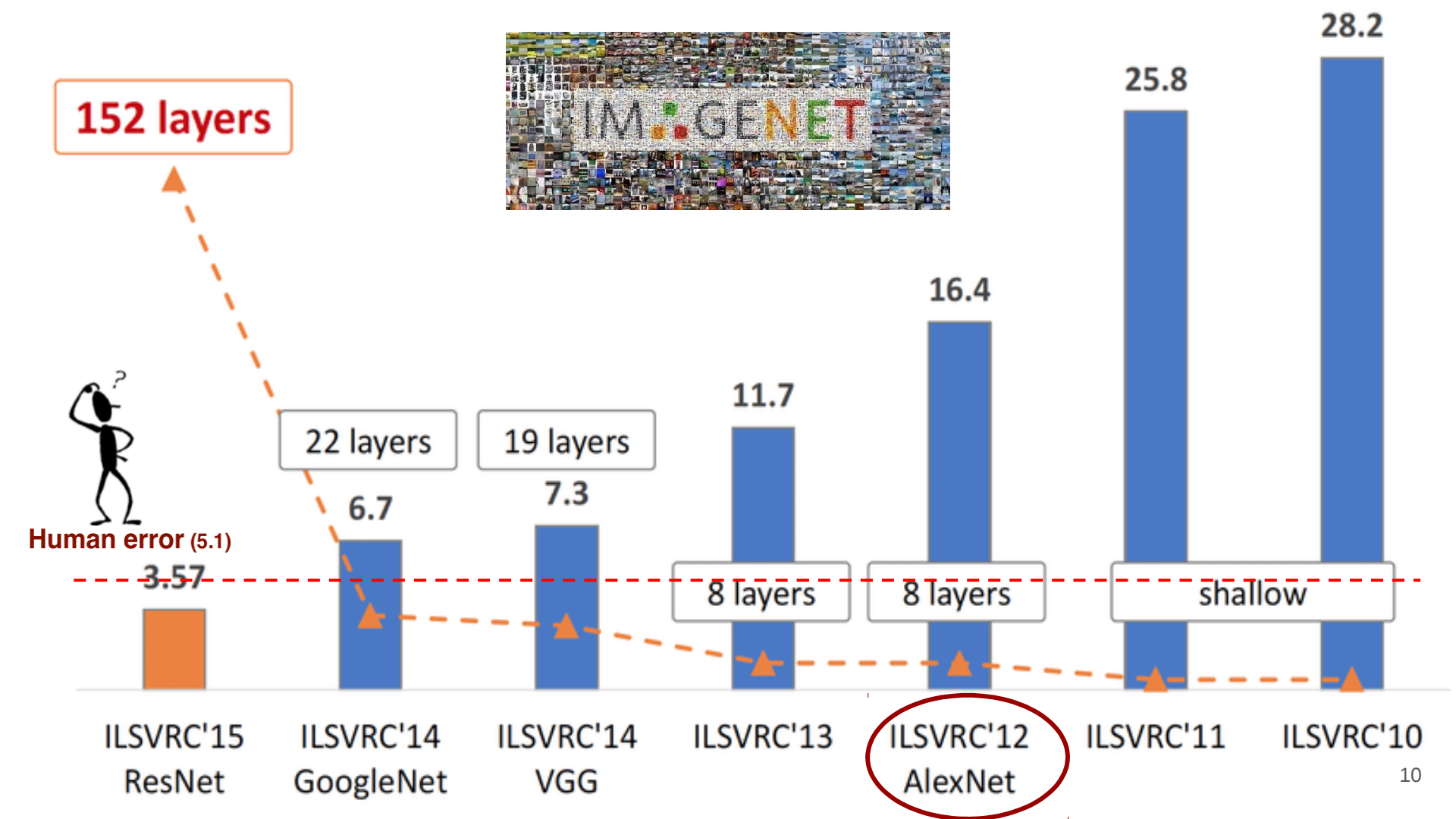
ILSVRC'14 VGG

ILSVRC'13

ILSVRC'12 AlexNet

ILSVRC'11

ILSVRC'10


152 layers
IMAGENET
28.2
25.8
16.4
11.7
22 layers
19 layers
6.7
7.3
Human error (5.1)
3.57
8 layers
8 layers
shallow
ILSVRC'15
ResNet
ILSVRC'14
GoogleNet
ILSVRC'14
VGG
ILSVRC'13
ILSVRC'12
AlexNet
ILSVRC'11
ILSVRC'10

152 layers

Human error (5.1)

3.57

22 layers

6.7

19 layers

7.3

11.7

8 layers

16.4

8 layers

25.8

shallow

28.2

ILSVRC'15 ResNet

ILSVRC'14 GoogleNet

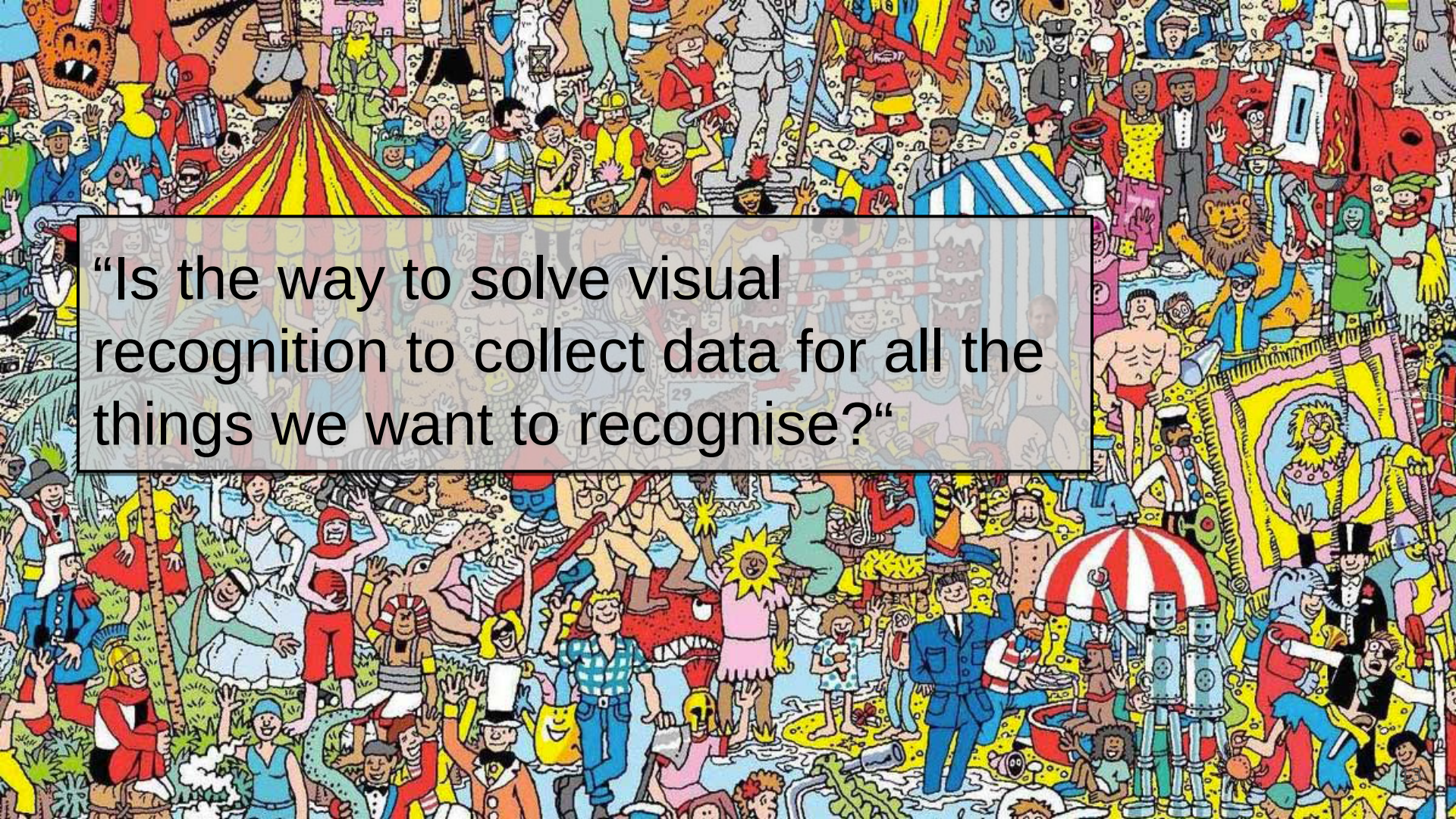
ILSVRC'14 VGG

ILSVRC'13

ILSVRC'12 AlexNet

ILSVRC'11

ILSVRC'10


“Is the way to solve visual recognition to collect data for all the things we want to recognise?“

[Russakovsky et al., 2015]

- 14M Images
- 22K Visual Concepts
- 10 years

[Russakovsky et al., 2015]

- 14M Images
- 22K Visual Concepts
- 10 years

[Biederman et al., 1987]

- 30K Nouns (1.4x)

[Russakovsky et al., 2015]

- 14M Images
- 22K Visual Concepts
- 10 years

[Biederman et al., 1987]

- 30K Nouns (1.4x)

[George A. Miller, 1995]

- Lexical database of English
- 175 979 Synset (8x)

[Russakovsky et al., 2015]

- 14M Images
- 22K Visual Concepts
- 10 years

[Biederman et al., 1987]

- 30K Nouns (1.4x)

[George A. Miller, 1995]

- Lexical database of English
- 175 979 Synset (8x)

[Facebook Inc., 2013]

- 300M images/day (21x)

[Russakovsky et al., 2015]

- 14M Images
- 22K Visual Concepts
- 10 years

[Biederman et al., 1987]

- 30K Nouns (1.4x)

[George A. Miller, 1995]

- Lexical database of English
- 175 979 Synset (8x)

[Facebook Inc., 2013]

- 300M images/day (21x)

[Bart Thomee et al., 2015]
8M visual concepts on the web (360x)

Structured Supervision

In this thesis, we propose methods that reduce the need for extensive human supervision by leveraging the structure in the visual world.

- Structure In The Outputs
- Structure In The Inputs
- Structure In The Models
- Leveraging Existing Models

Structure In the outputs

[On differentiating parameterized argmin and argmax problems with application to bi-level optimization.Stephen Gould, Basura Fernando, Anoop Cherian, Peter Anderson, **Rodrigo Santa Cruz**, Edison Guo. arXiv preprint arXiv:1607.05447, 2016.]

[DeepPermNet: Visual Permutation Learning. **Rodrigo Santa Cruz**, Basura Fernando, Anoop Cherian, Stephen Gould. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.]

[Visual Permutation Learning. **Rodrigo Santa Cruz**, Basura Fernando, Anoop Cherian, Stephen Gould. Pattern Analysis and Machine Intelligence (PAMI), 2018.]

Visual Labels are structured ...

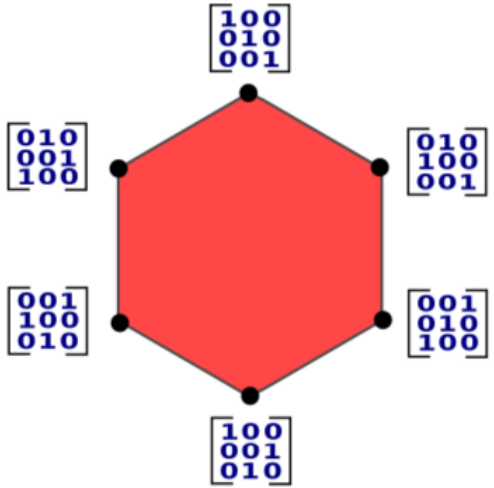


Image Ranking and Its Applications

Image Ranking and Its Applications

Subject: Smiling

Image Ranking and Its Applications

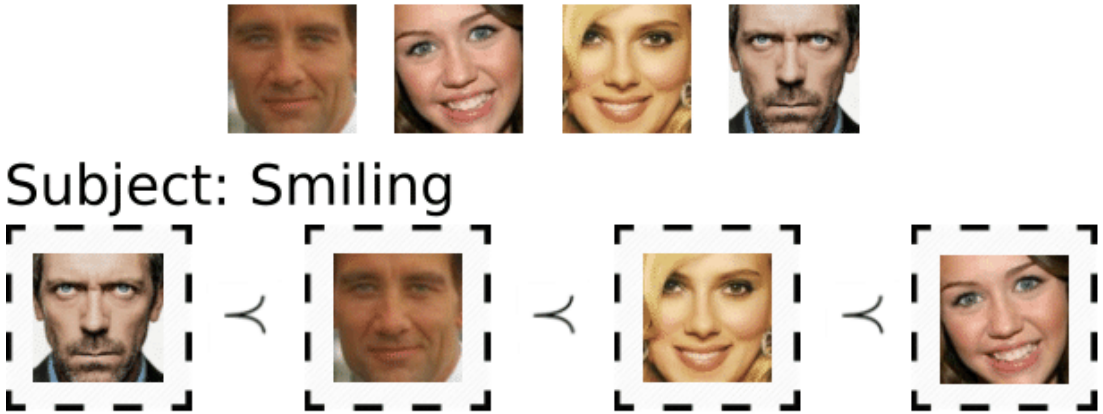


Image Ranking and Its Applications

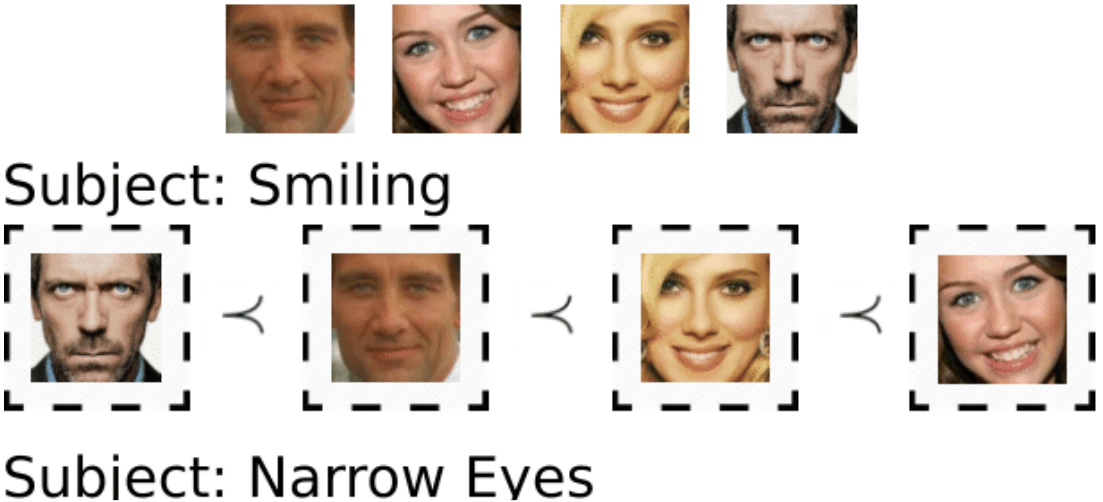


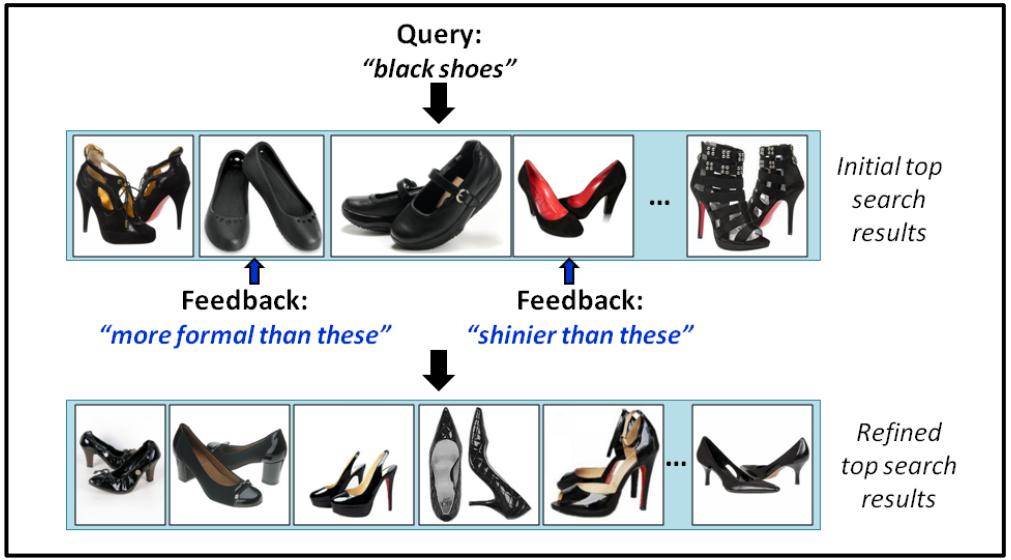
Image Ranking and Its Applications

Subject: Smiling

Subject: Narrow Eyes

Image Ranking Applications

Product Search Applications



Pearson Re-Identification

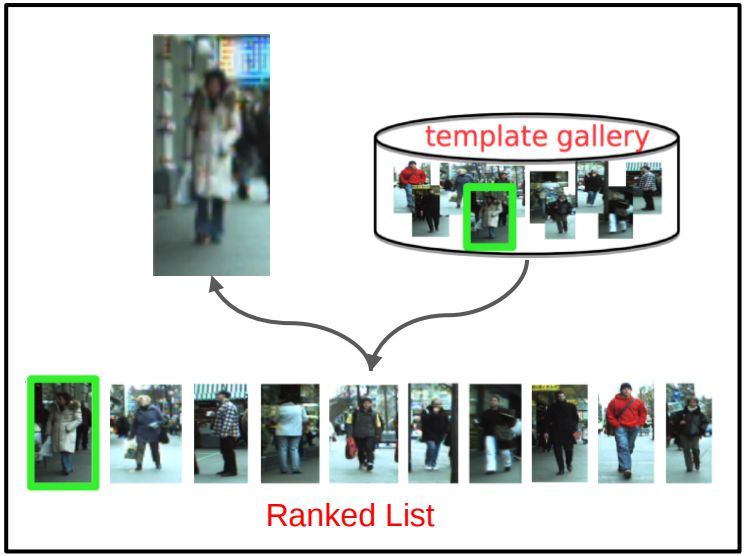
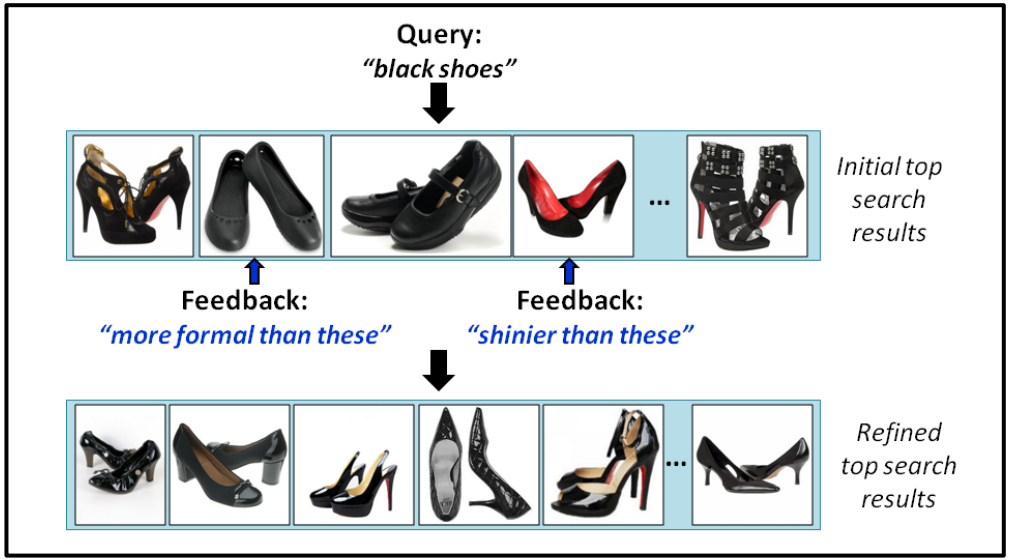
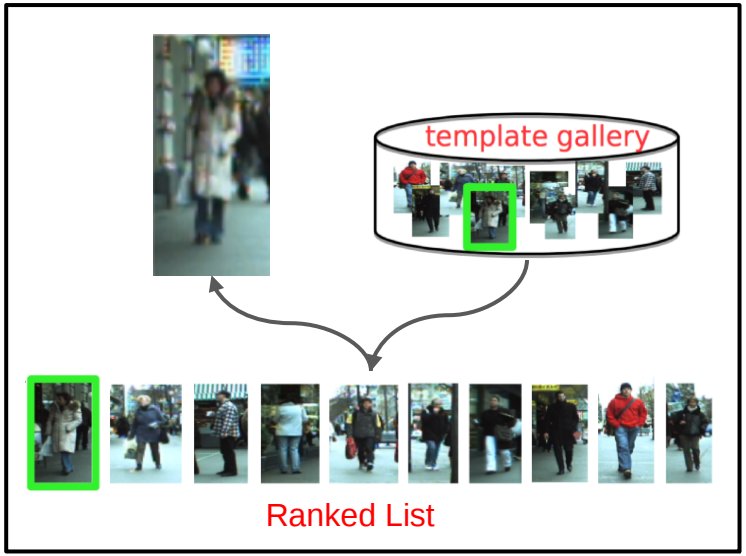


Image Ranking Applications

Product Search Applications



Pearson Re-Identification

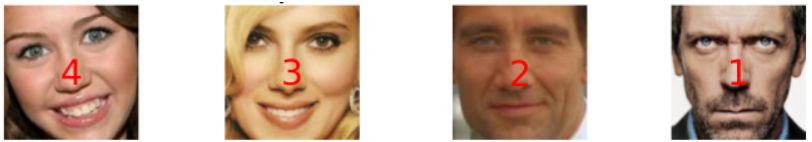


- Tasks in other fields can be reduced to this problem:
  - Computer graphics: Jigsaw puzzle
  - Biology: DNA and RNA modeling
  - Archeology: Re-assembling relics
- Archeology: Re-assembling relics
- **Computer Vision: Representation learning.**

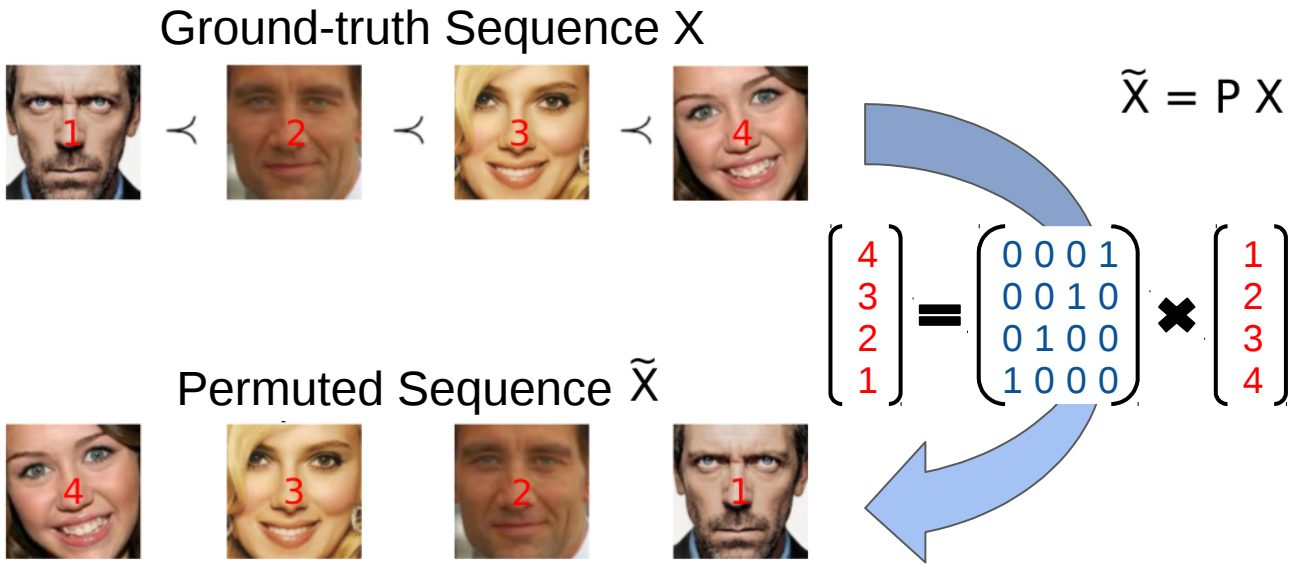
Visual Permutation Learning - Task

Ground-truth Sequence X

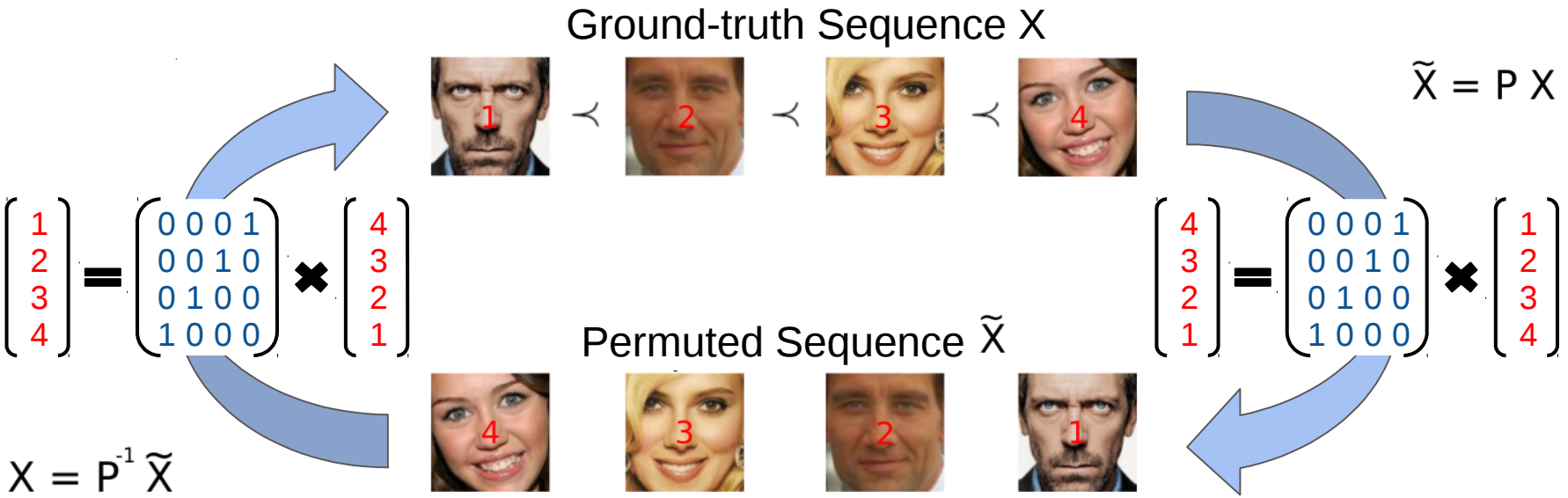
Permuted Sequence $\tilde{X}$



Visual Permutation Learning - Task



Visual Permutation Learning - Task



Visual Permutation Learning - Learning

Let us define a training set,

$$\mathcal{D} = \{(X, P) \mid X \in \mathcal{S}^c \text{ and } \forall P \in \mathcal{P}^l\}$$

We propose to learn a function that maps from fixed length image sequence to permutation matrices. Then our permutation learning problem can be described as,

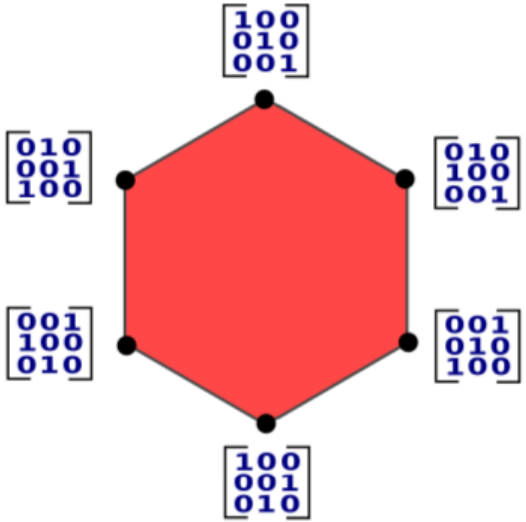
$$\underset{\theta}{\text{minimize}} \sum_{(X,P)\in\mathcal{D}} \Delta\left(P, f_\theta(\tilde{X})\right) + R(\theta)$$

Geometry of Permutation Matrices

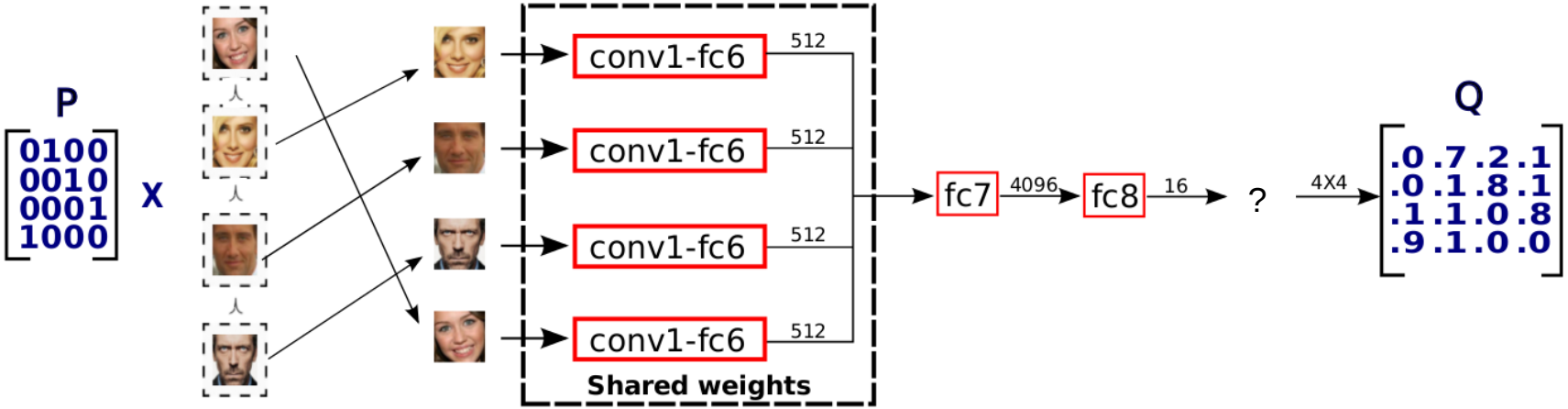
Then, we propose to approximate inference over permutation matrices to inference over their nearest convex-surrogate, the doubly stochastic matrices.

$$f_\theta \quad : \quad \mathcal{S}^c \quad \rightarrow \quad \mathcal{B}^l$$

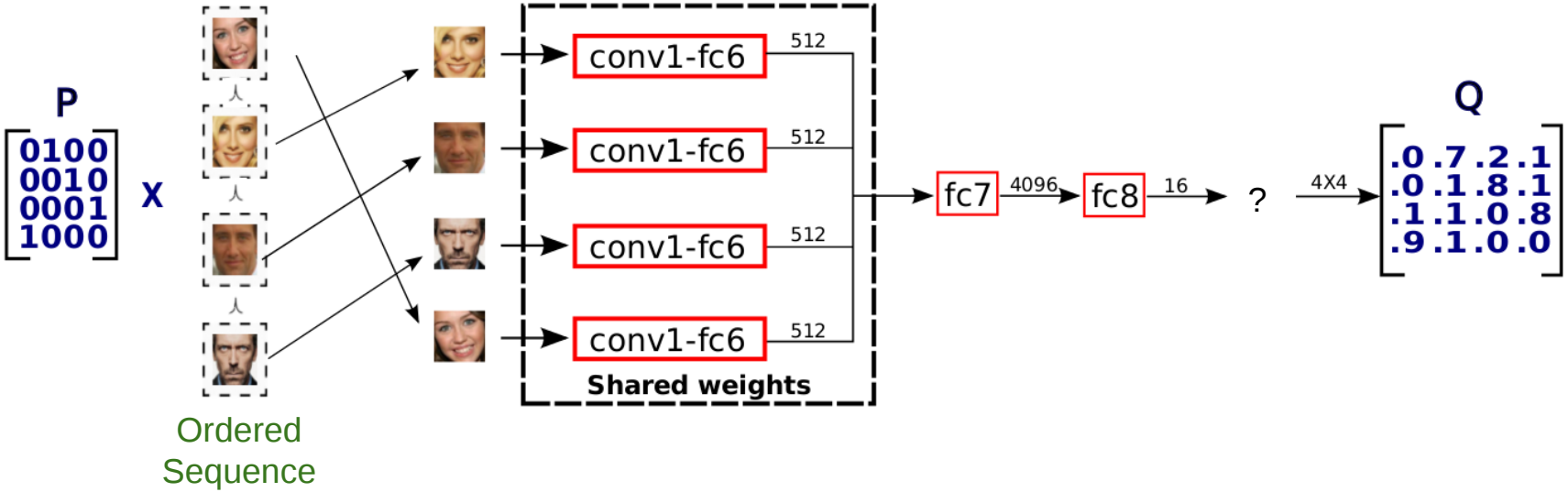
Learnable function Image Sequences doubly stochastic matrices



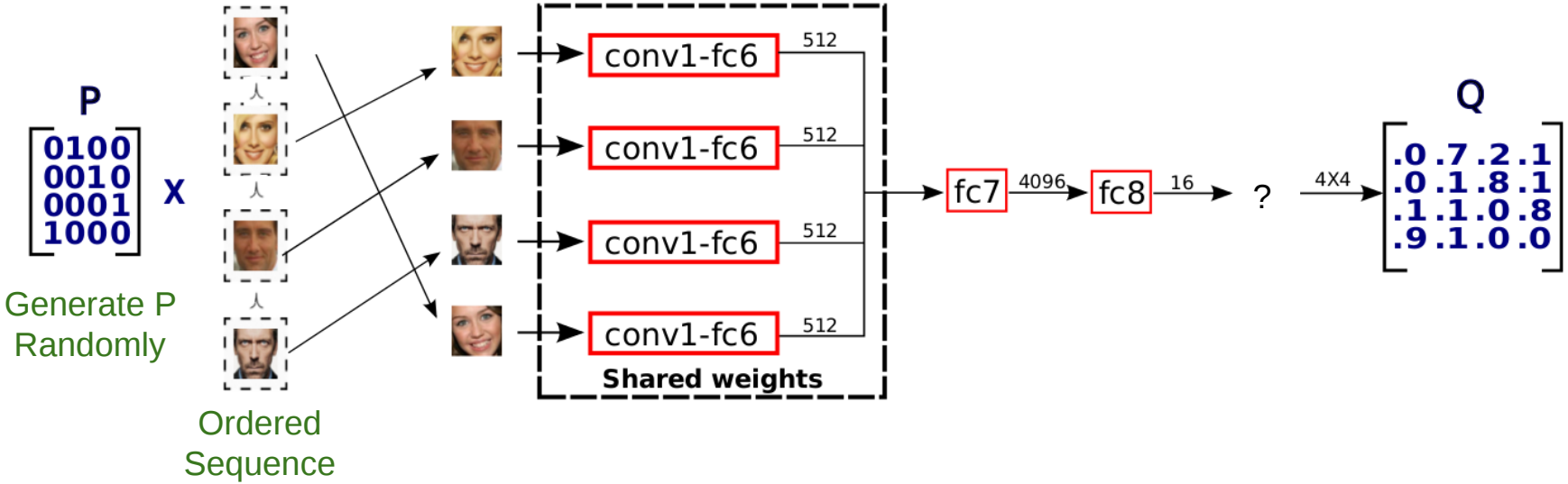
DeepPermNet - Model



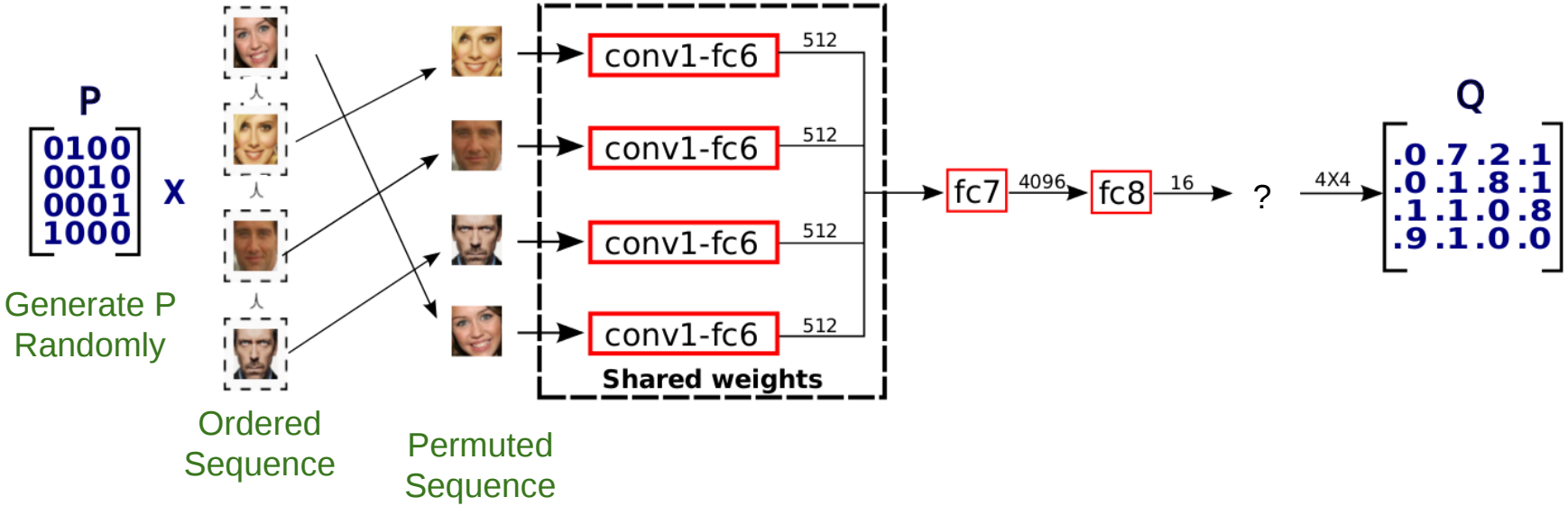
DeepPermNet - Model



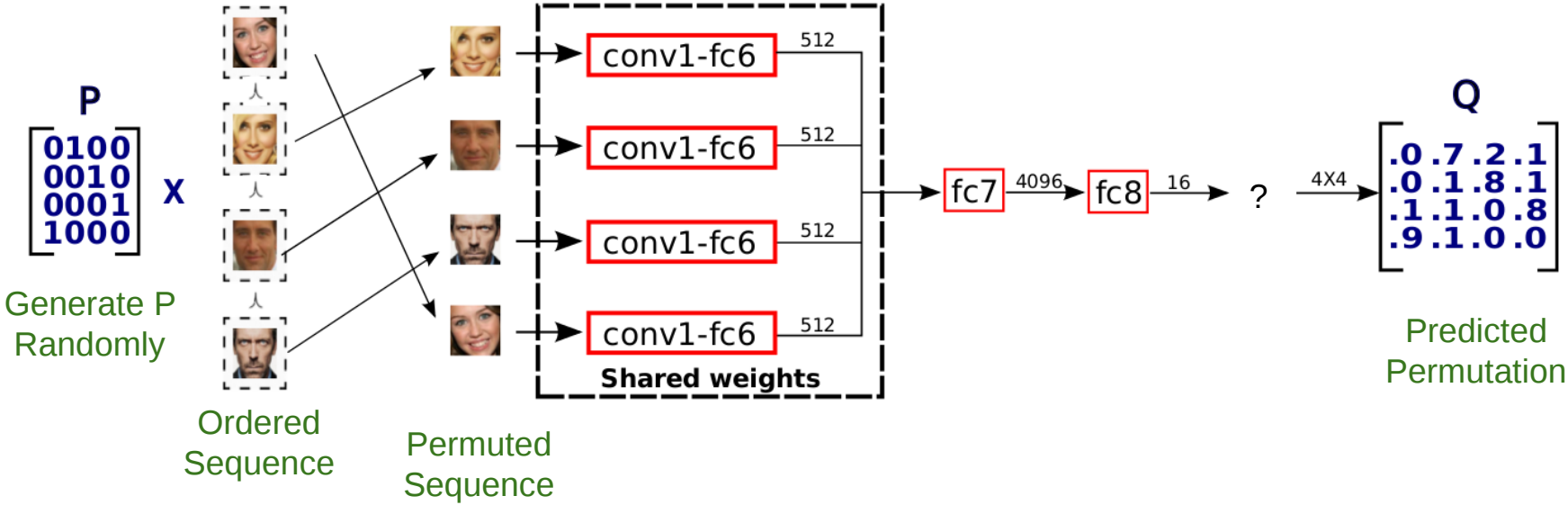
DeepPermNet - Model



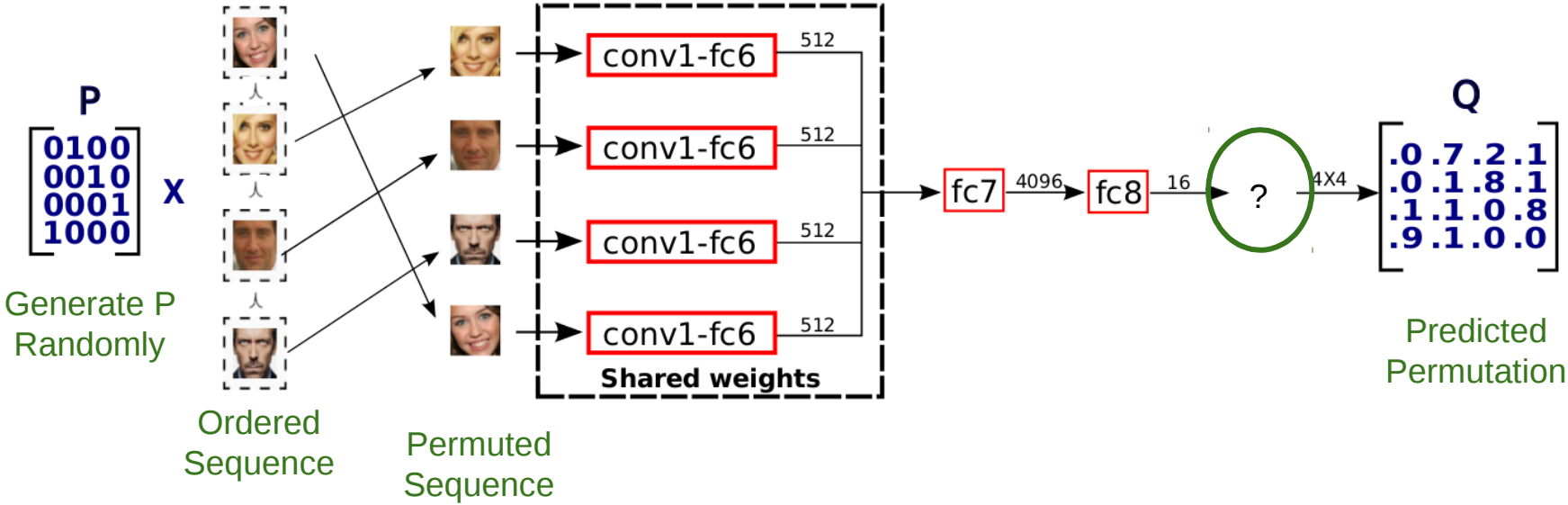
DeepPermNet - Model



DeepPermNet - Model



DeepPermNet - Model



Sinkhorn Layer

Sinkhorn's theorem: Any non-negative square matrix can be converted to a DSM by alternating between re-scaling its rows and columns to one.

Row normalization

$$R_{i,j}(Q) = \frac{Q_{i,j}}{\sum_{k=1}^{l} Q_{i,k}}; \quad C_{i,j}(Q) = \frac{Q_{i,j}}{\sum_{k=1}^{l} Q_{k,j}}$$

Column normalization

$$S^n(Q) = \begin{cases} Q, & \text{if } n = 0 \\ C\left(R\left(S^{n-1}(Q)\right)\right), & \text{otherwise.} \end{cases}$$

Gradient (Row normalization):

$$\frac{\partial \Delta}{\partial Q_{p,q}} = \sum_{j=1}^{l} \frac{\partial \Delta}{\partial R_{p,j}} \left[\frac{[\![j = q]\!]}{\sum_{k=1}^{l} Q_{p,k}} - \frac{Q_{p,j}}{\left(\sum_{k=1}^{l} Q_{p,k}\right)^2} \right]$$

Bi-level Optimization

$$
\begin{aligned}
\underset{\theta}{\text{minimize}} \quad & \sum_{(X,P)\in\mathcal{D}} \Delta\left(P, \hat{Q}\right) + R\left(\theta\right) \\
\text{subject to} \quad & \hat{Q} \in \underset{\hat{Q} \,\in\, \mathbb{R}_{+}^{n\times n}}{\text{argmin}} \quad \left\| \hat{Q} - f_\theta(\tilde{X}) \right\| \\
& \qquad\qquad \text{subject to} \quad \hat{Q}\,\mathbf{1} = \mathbf{1} \\
& \qquad\qquad\qquad\qquad\quad \hat{Q}^T\,\mathbf{1} = \mathbf{1}
\end{aligned}
$$

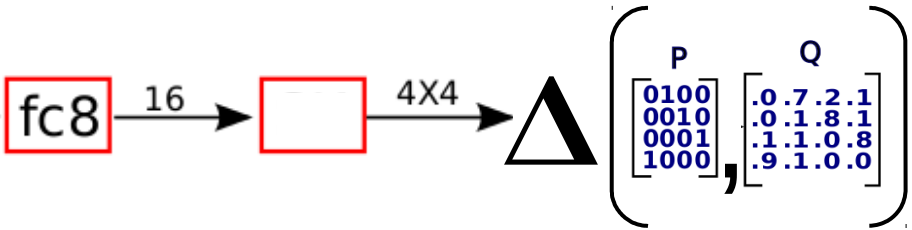
Bi-level Optimization

$$\underset{\theta}{\text{minimize}} \quad \sum_{(X,P)\in\mathcal{D}} \Delta\left(P, \hat{Q}\right) + R\left(\theta\right)$$

$$\text{subject to} \quad \hat{Q} \in \underset{\hat{Q} \in \mathbb{R}_{+}^{n\times n}}{\text{argmin}} \quad \left\|\hat{Q} - f_\theta(\tilde{X})\right\|$$

$$\text{subject to} \quad \hat{Q}\,\mathbf{1} = \mathbf{1}$$

$$\hat{Q}^T\,\mathbf{1} = \mathbf{1}$$



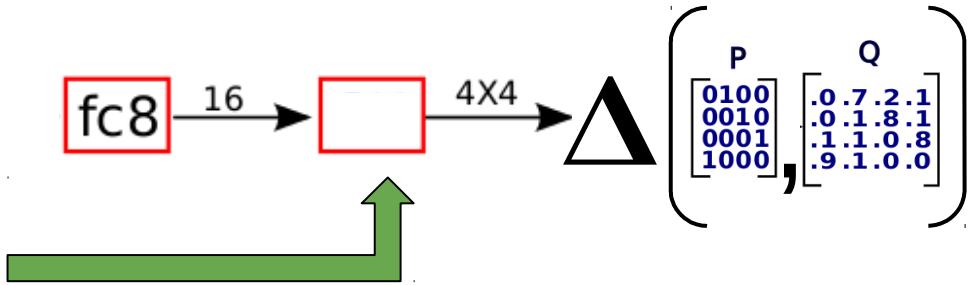
Bi-level Optimization

$$\underset{\theta}{\text{minimize}} \sum_{(X,P)\in\mathcal{D}} \Delta\left(P, \hat{Q}\right) + R(\theta)$$

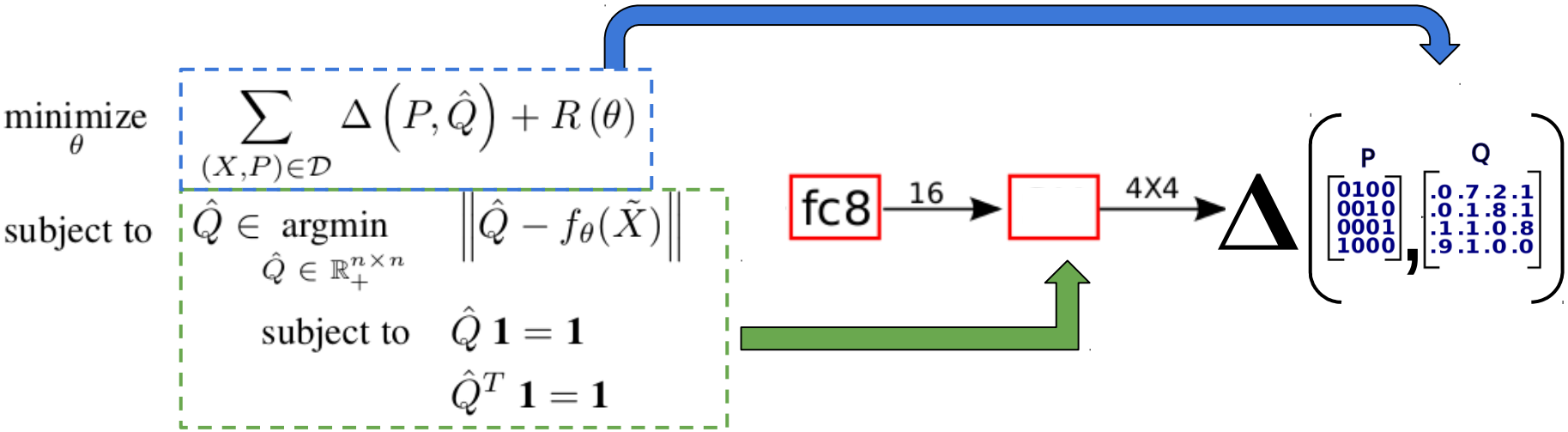
$$\text{subject to } \hat{Q} \in \underset{\hat{Q} \in \mathbb{R}_+^{n\times n}}{\text{argmin}} \left\|\hat{Q} - f_\theta(\tilde{X})\right\|$$

$$\text{subject to } \hat{Q}\,\mathbf{1} = \mathbf{1}$$

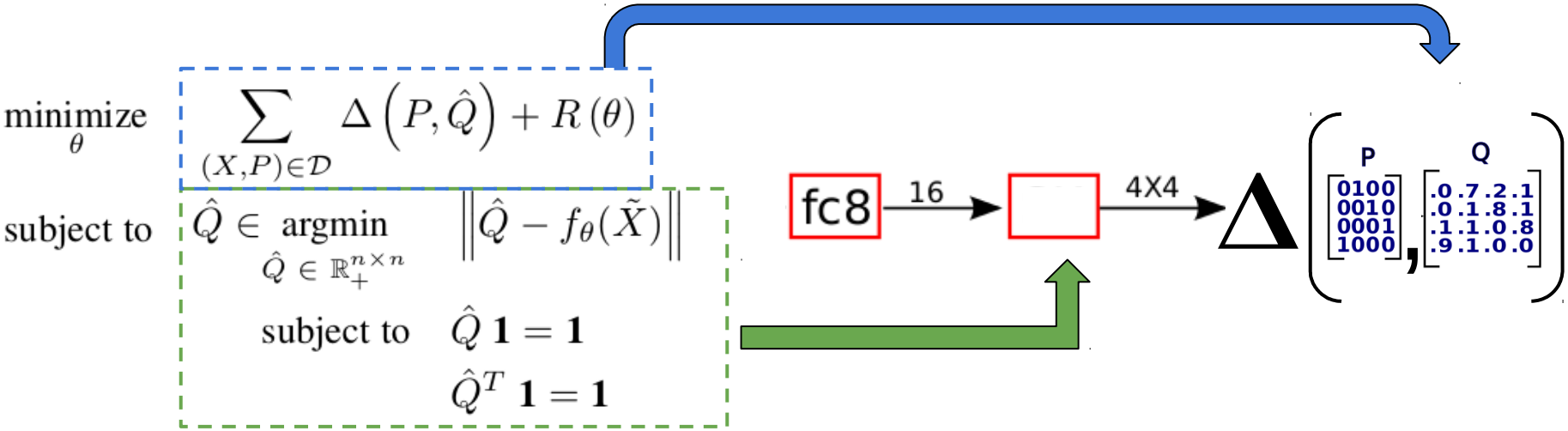
$$\hat{Q}^T\,\mathbf{1} = \mathbf{1}$$



Bi-level Optimization



Bi-level Optimization



We refer to "On differentiating parameterized argmin and argmax problems with application to bi-level optimization" by Gould et al. for a detailed explanation about computing gradients of argmin functions.

Visual Permutation Learning - Inference

We can recover the correctly ordered sequence from a permuted sequence by,

1 - Solving a approximation problem (or argmax rows/cols)

$$\hat{P} \in \underset{\hat{P}}{\operatorname{argmin}} \quad \left\| \hat{P} - Q \right\|_F$$
$$\text{subject to} \quad \hat{P} \cdot \mathbf{1} = \mathbf{1}$$
$$\mathbf{1}^T \cdot \hat{P} = \mathbf{1}$$
$$\hat{P} \in \{0, 1\}^{l \times l}$$

2 - Permuting the shuffled image sequence by the inverse permutation

$$X = \hat{P}^T \tilde{X}$$

Visual Permutation Learning - Recap

- Given a set of ordered images Sc according to c, we build a data set D as,

$$\mathcal{D} = \{(X, P) \mid X \in \mathcal{S}^c \text{ and } \forall P \in \mathcal{P}^l\}$$

- Using D, we learn a function (CNN) which maps shuffled image sequences to its DSM matrix employing the sinkhorn layer or bi-level optimization.

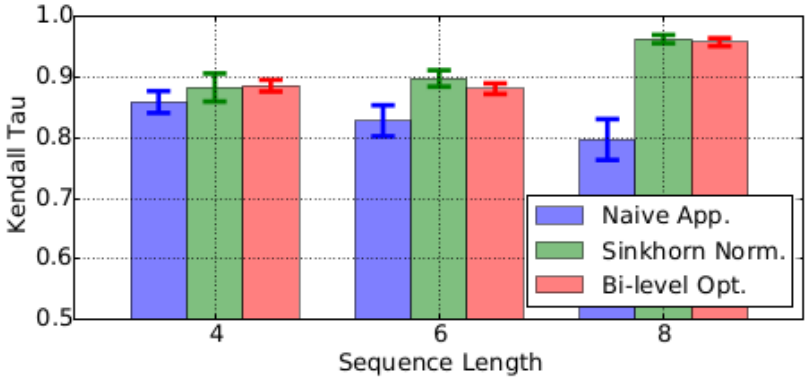
$$f_\theta : \mathcal{S}^c \rightarrow \mathcal{B}^l$$

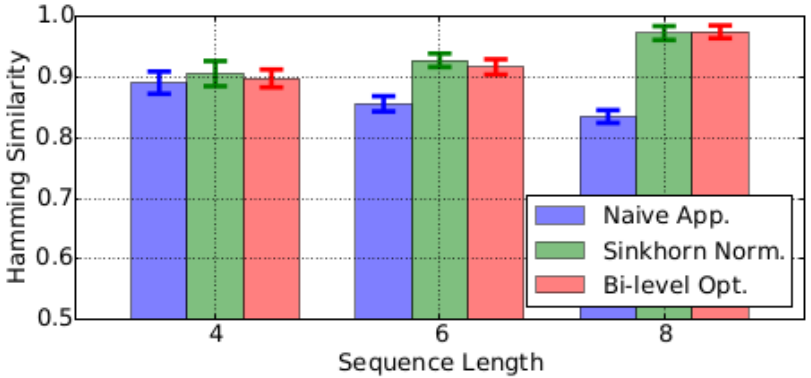
- During test time, we receive a shuffled image sequence and reorder it according to c by doing,

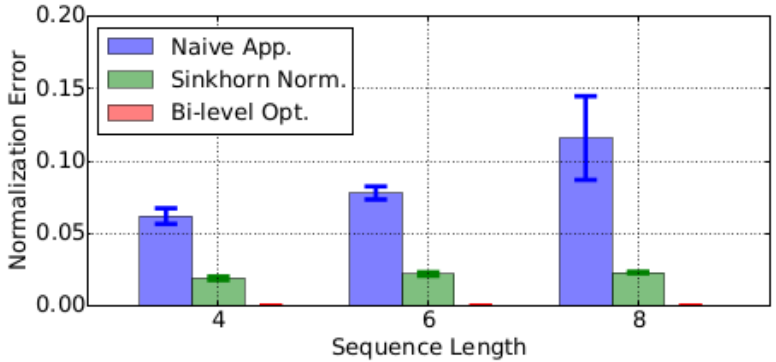
$$\tilde{X} \rightarrow f_\theta(\cdot) \rightarrow \text{Infer P} \rightarrow X = P^T \tilde{X}$$

Experiments - Permutation Prediction

Unpermute 20K shuffled sequences:

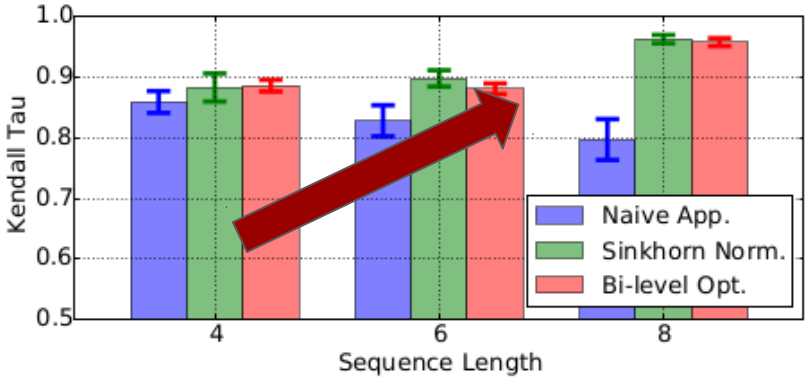


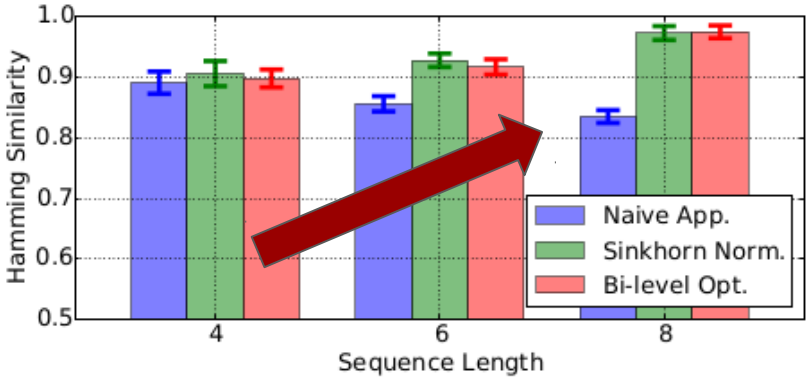


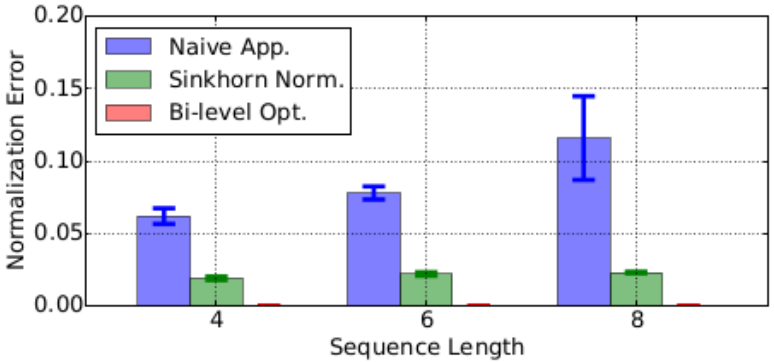


Experiments - Permutation Prediction

Unpermute 20K shuffled sequences:

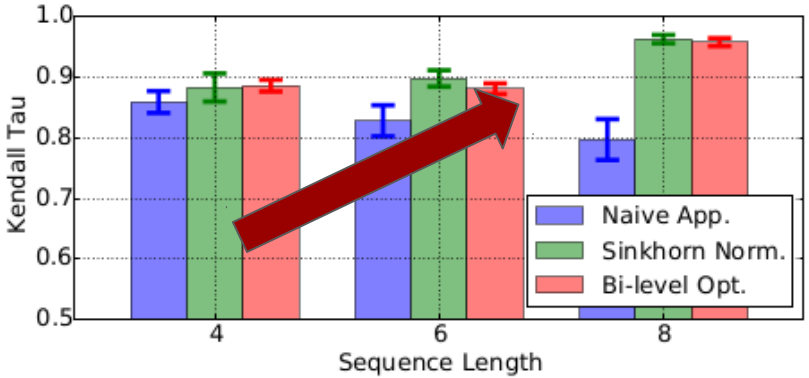


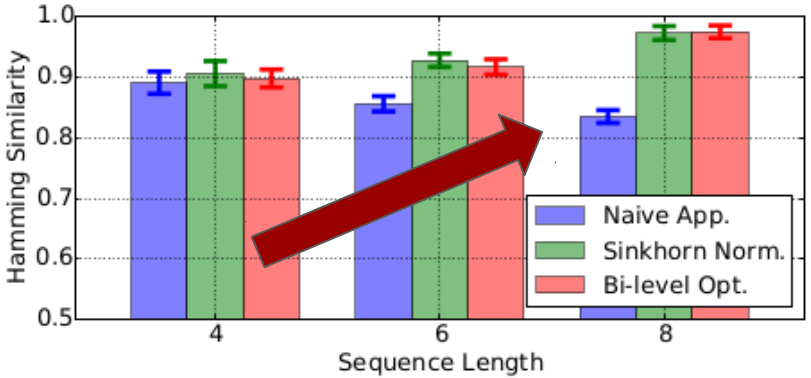


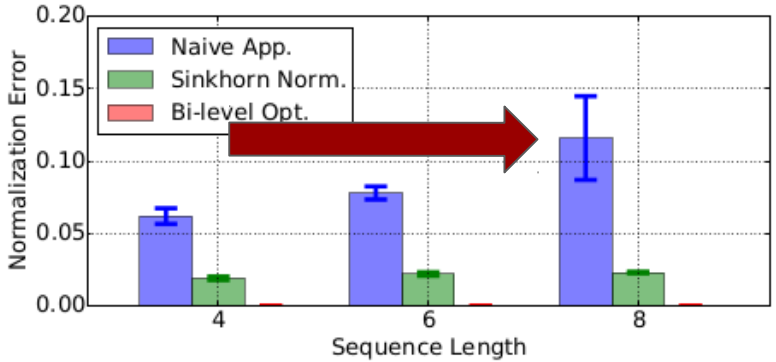


Experiments - Permutation Prediction

Unpermute 20K shuffled sequences:







Experiments - Relative Attributes

TABLE 1
Evaluating the proposed model applied to the relative attributes task on the Public Figures Dataset. We report the pairwise accuracy as well as its mean across the attributes.

Method	Lips	Eyebrows	Chubby	Male	Eyes	Nose	Face	Smiling	Forehead	White	Young	Mean
Parikh and Grauman [59]	79.17	79.87	76.27	81.80	81.67	77.40	82.33	79.90	87.60	76.97	83.20	80.56
Li et al. [46]	81.87	81.84	79.97	85.33	83.15	80.43	86.31	83.36	88.83	82.59	84.41	83.37
Yu and Grauman [82]	90.43	89.83	87.37	91.77	91.40	89.07	86.70	87.00	94.00	87.43	91.87	89.72
Souri et al. [71]	93.62	94.53	92.32	95.50	93.19	94.24	94.76	95.36	97.28	94.60	94.33	94.52
DeepPermNet (Sinkhorn Norm.)	**99.55**	**97.21**	97.66	**99.44**	96.54	**96.21**	99.11	97.88	**99.00**	**97.99**	**99.00**	**98.14**
DeepPermNet (Bi-level Opt.)	99.53	96.65	**98.54**	98.99	**97.21**	94.72	**99.44**	**98.55**	98.77	95.66	98.77	97.89

TABLE 2
Evaluating the proposed model applied to the relative attributes task on the OSR dataset. We report the pairwise accuracy as well as its mean across the attributes.

Method	Depth-Close	Diagonal-Plane	Natural	Open	Perspective	Size-Large	Mean
Parikh and Grauman [59]	87.53	86.5	95.03	90.77	86.73	86.23	88.80
Li et al. [46]	89.54	89.34	95.24	92.39	87.58	88.34	90.41
Yu and Grauman [82]	90.47	92.43	95.7	94.1	90.43	91.1	92.37
Singh and Lee [69]	96.1	97.64	98.89	97.2	96.31	95.98	97.02
Souri et al. [71]	97.65	98.43	**99.4**	97.44	96.88	96.79	97.77
DeepPermNet (Sinkhorn Norm.)	96.09	94.53	97.21	96.65	96.46	98.77	96.62
DeepPermNet (Bi-level Opt.)	97.99	98.21	97.76	97.10	97.21	96.65	97.49
DeepPermNet (Sinkhorn Norm. + VGG16)	96.87	97.99	96.87	**99.79**	**99.82**	**99.55**	**98.48**
DeepPermNet (Bi-level Opt. + VGG16)	**98.12**	**99.92**	98.13	97.78	98.72	97.87	98.42

Experiments - Relative Attributes

TABLE 1
Evaluating the proposed model applied to the relative attributes task on the Public Figures Dataset. We report the pairwise accuracy as well as its mean across the attributes.

Method	Lips	Eyebrows	Chubby	Male	Eyes	Nose	Face	Smiling	Forehead	White	Young	Mean
Parikh and Grauman [59]	79.17	79.87	76.27	81.80	81.67	77.40	82.33	79.90	87.60	76.97	83.20	80.56
Li et al. [46]	81.87	81.84	79.97	85.33	83.15	80.43	86.31	83.36	88.83	82.59	84.41	83.37
Yu and Grauman [82]	90.43	89.83	87.37	91.77	91.40	89.07	86.70	87.00	94.00	87.43	91.87	89.72
Souri et al. [71] VGG	93.62	94.53	92.32	95.50	93.19	94.24	94.76	95.36	97.28	94.60	94.33	94.52
DeepPermNet (Sinkhorn Norm.)	**99.55**	**97.21**	97.66	**99.44**	96.54	**96.21**	99.11	97.88	**99.00**	**97.99**	**99.00**	**98.14**
DeepPermNet (Bi-level Opt.)	99.53	96.65	**98.54**	98.99	**97.21**	94.72	**99.44**	**98.55**	98.77	95.66	98.77	97.89

TABLE 2
Evaluating the proposed model applied to the relative attributes task on the OSR dataset. We report the pairwise accuracy as well as its mean across the attributes.

Method	Depth-Close	Diagonal-Plane	Natural	Open	Perspective	Size-Large	Mean
Parikh and Grauman [59]	87.53	86.5	95.03	90.77	86.73	86.23	88.80
Li et al. [46]	89.54	89.34	95.24	92.39	87.58	88.34	90.41
Yu and Grauman [82]	90.47	92.43	95.7	94.1	90.43	91.1	92.37
Singh and Lee [69]	96.1	97.64	98.89	97.2	96.31	95.98	97.02
Souri et al. [71]	97.65	98.43	**99.4**	97.44	96.88	96.79	97.77
DeepPermNet (Sinkhorn Norm.)	96.09	94.53	97.21	96.65	96.46	98.77	96.62
DeepPermNet (Bi-level Opt.)	97.99	98.21	97.76	97.10	97.21	96.65	97.49
DeepPermNet (Sinkhorn Norm. + VGG16)	96.87	97.99	96.87	**99.79**	**99.82**	**99.55**	**98.48**
DeepPermNet (Bi-level Opt. + VGG16)	**98.12**	**99.92**	98.13	97.78	98.72	97.87	98.42

Experiments - Relative Attributes

TABLE 1
Evaluating the proposed model applied to the relative attributes task on the Public Figures Dataset. We report the pairwise accuracy as well as its mean across the attributes.

Method	Lips	Eyebrows	Chubby	Male	Eyes	Nose	Face	Smiling	Forehead	White	Young	Mean
Parikh and Grauman [59]	79.17	79.87	76.27	81.80	81.67	77.40	82.33	79.90	87.60	76.97	83.20	80.56
Li et al. [46]	81.87	81.84	79.97	85.33	83.15	80.43	86.31	83.36	88.83	82.59	84.41	83.37
Yu and Grauman [82]	90.43	89.83	87.37	91.77	91.40	89.07	86.70	87.00	94.00	87.43	91.87	89.72
Souri et al. [71] VGG	93.62	94.53	92.32	95.50	93.19	94.24	94.76	95.36	97.28	94.60	94.33	94.52
DeepPermNet (Sinkhorn Norm.)	**99.55**	**97.21**	97.66	**99.44**	96.54	**96.21**	99.11	97.88	**99.00**	**97.99**	**99.00**	**98.14**
DeepPermNet (Bi-level Opt.)	99.53	96.65	**98.54**	98.99	**97.21**	94.72	**99.44**	**98.55**	98.77	95.66	98.77	97.89

TABLE 2
Evaluating the proposed model applied to the relative attributes task on the OSR dataset. We report the pairwise accuracy as well as its mean across the attributes.

Method	Depth-Close	Diagonal-Plane	Natural	Open	Perspective	Size-Large	Mean
Parikh and Grauman [59]	87.53	86.5	95.03	90.77	86.73	86.23	88.80
Li et al. [46]	89.54	89.34	95.24	92.39	87.58	88.34	90.41
Yu and Grauman [82]	90.47	92.43	95.7	94.1	90.43	91.1	92.37
Singh and Lee [69]	96.1	97.64	98.89	97.2	96.31	95.98	97.02
Souri et al. [71]	97.65	98.43	**99.4**	97.44	96.88	96.79	97.77
DeepPermNet (Sinkhorn Norm.)	96.09	94.53	97.21	96.65	96.46	98.77	96.62
DeepPermNet (Bi-level Opt.)	97.99	98.21	97.76	97.10	97.21	96.65	97.49
DeepPermNet (Sinkhorn Norm. + VGG16)	96.87	97.99	96.87	**99.79**	**99.82**	**99.55**	**98.48**
DeepPermNet (Bi-level Opt. + VGG16)	**98.12**	**99.92**	98.13	97.78	98.72	97.87	98.42

Experiments - Relative Attributes

TABLE 1
Evaluating the proposed model applied to the relative attributes task on the Public Figures Dataset. We report the pairwise accuracy as well as its mean across the attributes.

Method	Lips	Eyebrows	Chubby	Male	Eyes	Nose	Face	Smiling	Forehead	White	Young	Mean
Parikh and Grauman [59]	79.17	79.87	76.27	81.80	81.67	77.40	82.33	79.90	87.60	76.97	83.20	80.56
Li et al. [46]	81.87	81.84	79.97	85.33	83.15	80.43	86.31	83.36	88.83	82.59	84.41	83.37
Yu and Grauman [82]	90.43	89.83	87.37	91.77	91.40	89.07	86.70	87.00	94.00	87.43	91.87	89.72
Souri et al. [71] VGG	93.62	94.53	92.32	95.50	93.19	94.24	94.76	95.36	97.28	94.60	94.33	94.52
DeepPermNet (Sinkhorn Norm.)	**99.55**	**97.21**	97.66	**99.44**	96.54	**96.21**	99.11	97.88	**99.00**	**97.99**	**99.00**	**98.14**
DeepPermNet (Bi-level Opt.)	99.53	96.65	**98.54**	98.99	**97.21**	94.72	**99.44**	**98.55**	98.77	95.66	98.77	97.89
Δ =	5.93		7.22				4.68				4.67	3.62

TABLE 2
Evaluating the proposed model applied to the relative attributes task on the OSR dataset. We report the pairwise accuracy as well as its mean across the attributes.

Method	Depth-Close	Diagonal-Plane	Natural	Open	Perspective	Size-Large	Mean
Parikh and Grauman [59]	87.53	86.5	95.03	90.77	86.73	86.23	88.80
Li et al. [46]	89.54	89.34	95.24	92.39	87.58	88.34	90.41
Yu and Grauman [82]	90.47	92.43	95.7	94.1	90.43	91.1	92.37
Singh and Lee [69]	96.1	97.64	98.89	97.2	96.31	95.98	97.02
Souri et al. [71]	97.65	98.43	**99.4**	97.44	96.88	96.79	97.77
DeepPermNet (Sinkhorn Norm.)	96.09	94.53	97.21	96.65	96.46	98.77	96.62
DeepPermNet (Bi-level Opt.)	97.99	98.21	97.76	97.10	97.21	96.65	97.49
DeepPermNet (Sinkhorn Norm. + VGG16)	96.87	97.99	96.87	**99.79**	**99.82**	**99.55**	**98.48**
DeepPermNet (Bi-level Opt. + VGG16)	**98.12**	**99.92**	98.13	97.78	98.72	97.87	98.42

Experiments - Relative Attributes

TABLE 1
Evaluating the proposed model applied to the relative attributes task on the Public Figures Dataset. We report the pairwise accuracy as well as its mean across the attributes.

Method	Lips	Eyebrows	Chubby	Male	Eyes	Nose	Face	Smiling	Forehead	White	Young	Mean
Parikh and Grauman [59]	79.17	79.87	76.27	81.80	81.67	77.40	82.33	79.90	87.60	76.97	83.20	80.56
Li et al. [46]	81.87	81.84	79.97	85.33	83.15	80.43	86.31	83.36	88.83	82.59	84.41	83.37
Yu and Grauman [82]	90.43	89.83	87.37	91.77	91.40	89.07	86.70	87.00	94.00	87.43	91.87	89.72
Souri et al. [71] VGG	93.62	94.53	92.32	95.50	93.19	94.24	94.76	95.36	97.28	94.60	94.33	94.52
DeepPermNet (Sinkhorn Norm.)	**99.55**	**97.21**	97.66	**99.44**	96.54	**96.21**	99.11	97.88	**99.00**	**97.99**	**99.00**	**98.14**
DeepPermNet (Bi-level Opt.)	99.53	96.65	**98.54**	98.99	**97.21**	94.72	**99.44**	**98.55**	98.77	95.66	98.77	97.89
Δ =	5.93		7.22				4.68				4.67	3.62

TABLE 2
Evaluating the proposed model applied to the relative attributes task on the OSR dataset. We report the pairwise accuracy as well as its mean across the attributes.

Method	Depth-Close	Diagonal-Plane	Natural	Open	Perspective	Size-Large	Mean
Parikh and Grauman [59]	87.53	86.5	95.03	90.77	86.73	86.23	88.80
Li et al. [46]	89.54	89.34	95.24	92.39	87.58	88.34	90.41
Yu and Grauman [82]	90.47	92.43	95.7	94.1	90.43	91.1	92.37
Singh and Lee [69]	96.1	97.64	98.89	97.2	96.31	95.98	97.02
Souri et al. [71] VGG	97.65	98.43	**99.4**	97.44	96.88	96.79	97.77
DeepPermNet (Sinkhorn Norm.)	96.09	94.53	97.21	96.65	96.46	98.77	96.62
DeepPermNet (Bi-level Opt.)	97.99	98.21	97.76	97.10	97.21	96.65	97.49
DeepPermNet (Sinkhorn Norm. + VGG16)	96.87	97.99	96.87	**99.79**	**99.82**	**99.55**	**98.48**
DeepPermNet (Bi-level Opt. + VGG16)	**98.12**	**99.92**	98.13	97.78	98.72	97.87	98.42

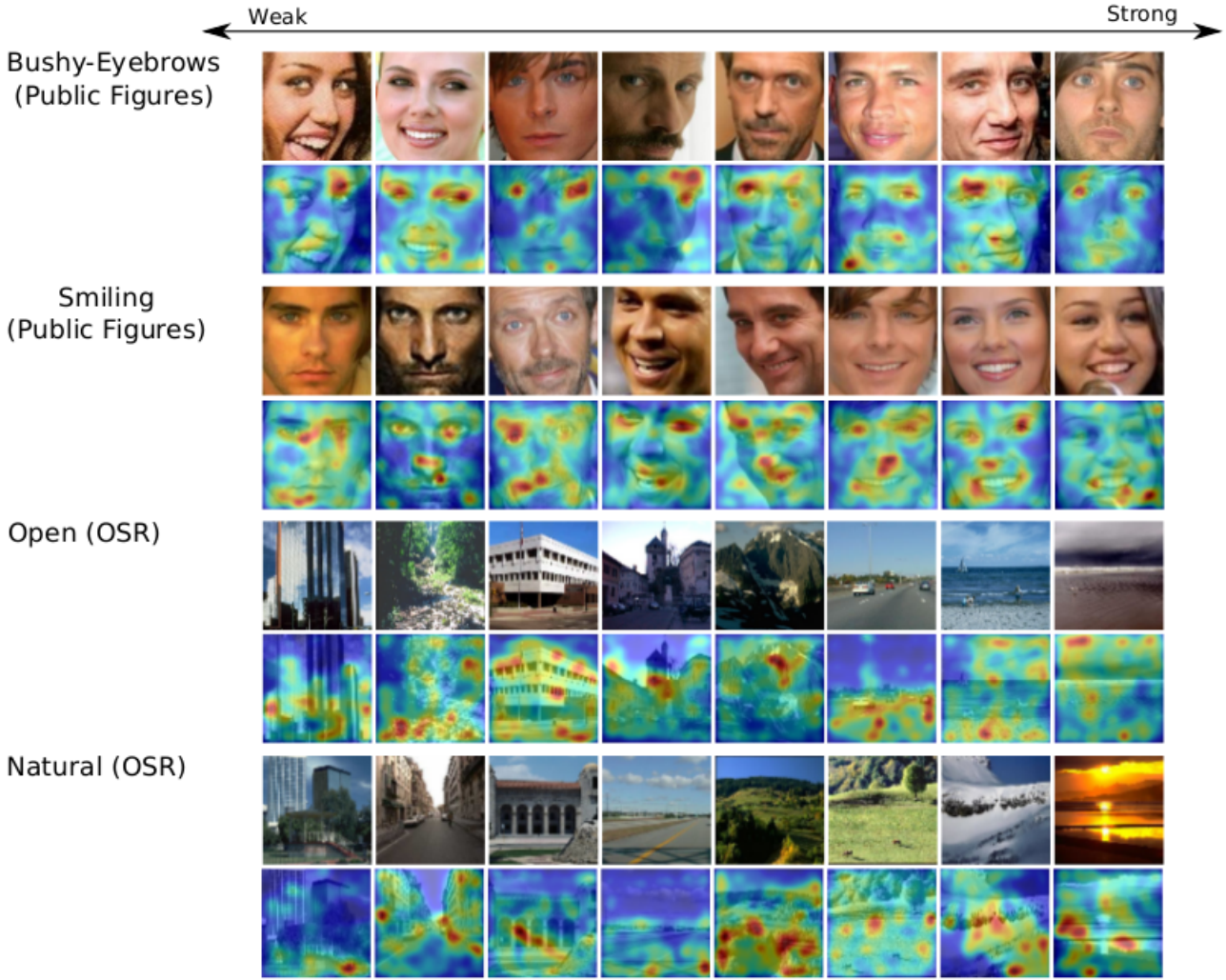
Experiments - Relative Attributes

TABLE 1
Evaluating the proposed model applied to the relative attributes task on the Public Figures Dataset. We report the pairwise accuracy as well as its mean across the attributes.

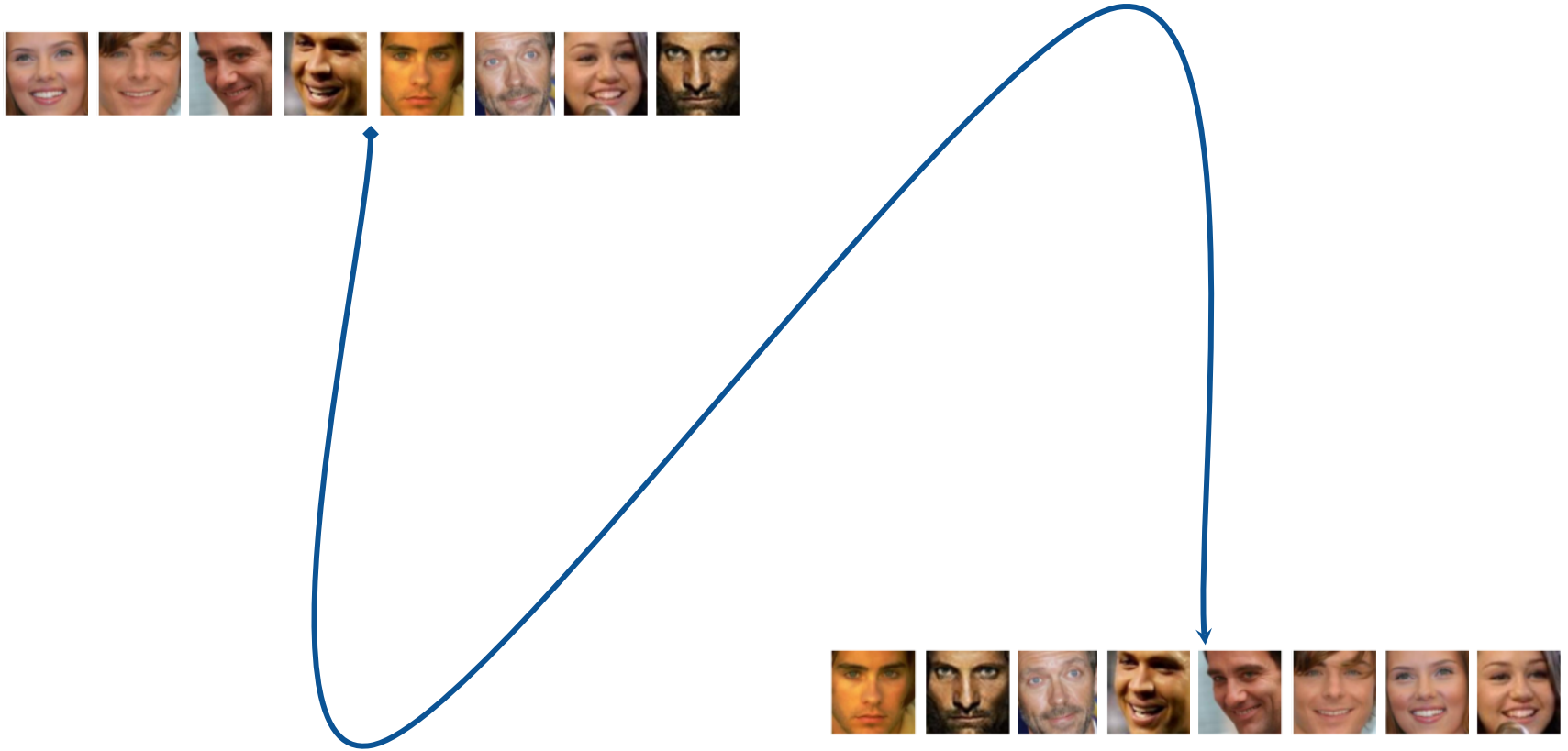
Method	Lips	Eyebrows	Chubby	Male	Eyes	Nose	Face	Smiling	Forehead	White	Young	Mean
Parikh and Grauman [59]	79.17	79.87	76.27	81.80	81.67	77.40	82.33	79.90	87.60	76.97	83.20	80.56
Li et al. [46]	81.87	81.84	79.97	85.33	83.15	80.43	86.31	83.36	88.83	82.59	84.41	83.37
Yu and Grauman [82]	90.43	89.83	87.37	91.77	91.40	89.07	86.70	87.00	94.00	87.43	91.87	89.72
Souri et al. [71] VGG	93.62	94.53	92.32	95.50	93.19	94.24	94.76	95.36	97.28	94.60	94.33	94.52
DeepPermNet (Sinkhorn Norm.)	**99.55**	**97.21**	97.66	**99.44**	96.54	**96.21**	99.11	97.88	**99.00**	**97.99**	**99.00**	**98.14**
DeepPermNet (Bi-level Opt.)	99.53	96.65	**98.54**	98.99	**97.21**	94.72	**99.44**	**98.55**	98.77	95.66	98.77	97.89
Δ =	5.93		7.22				4.68				4.67	3.62

TABLE 2
Evaluating the proposed model applied to the relative attributes task on the OSR dataset. We report the pairwise accuracy as well as its mean across the attributes.

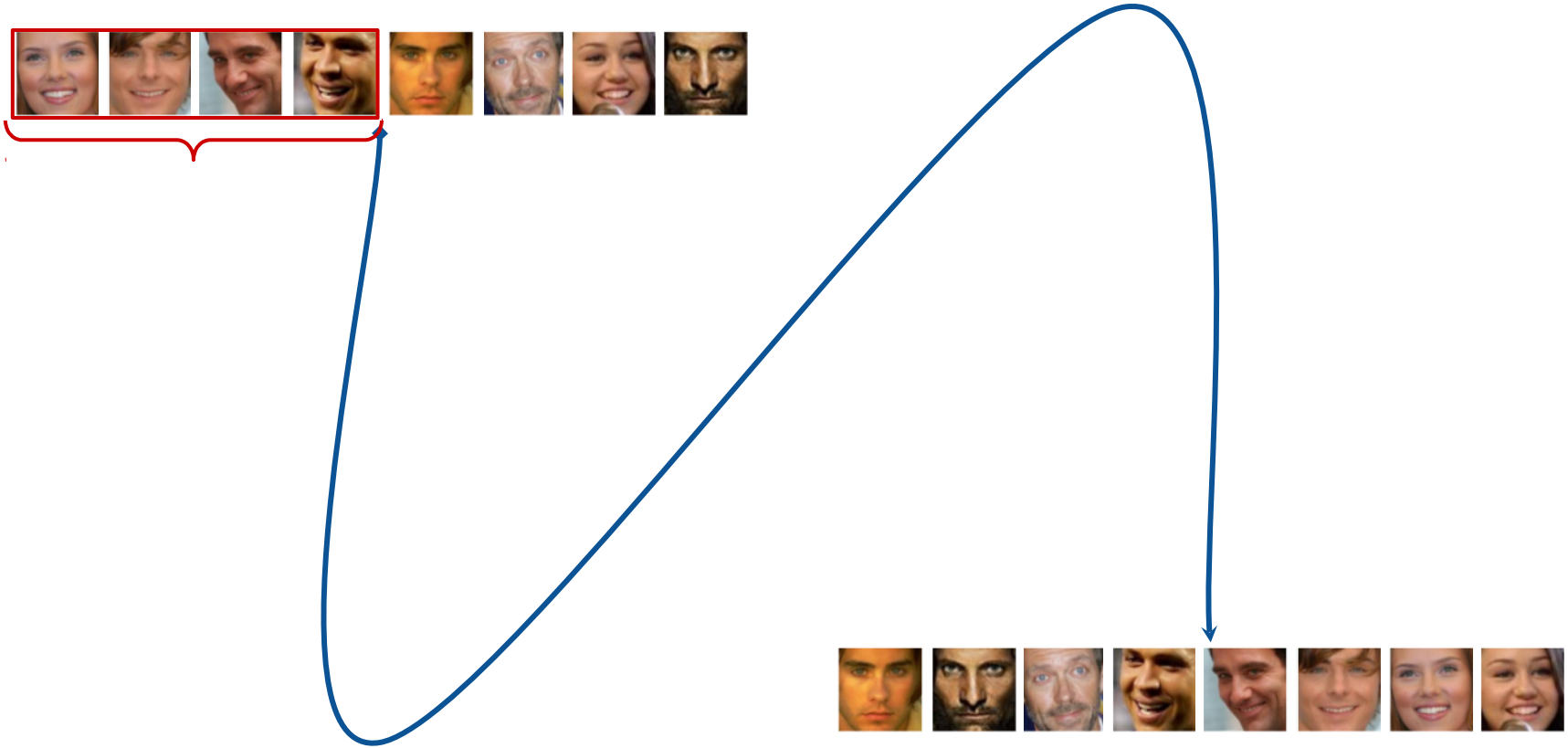
Method	Depth-Close	Diagonal-Plane	Natural	Open	Perspective	Size-Large	Mean
Parikh and Grauman [59]	87.53	86.5	95.03	90.77	86.73	86.23	88.80
Li et al. [46]	89.54	89.34	95.24	92.39	87.58	88.34	90.41
Yu and Grauman [82]	90.47	92.43	95.7	94.1	90.43	91.1	92.37
Singh and Lee [69]	96.1	97.64	98.89	97.2	96.31	95.98	97.02
Souri et al. [71] VGG	97.65	98.43	**99.4**	97.44	96.88	96.79	97.77
DeepPermNet (Sinkhorn Norm.)	96.09	94.53	97.21	96.65	96.46	98.77	96.62
DeepPermNet (Bi-level Opt.)	97.99	98.21	97.76	97.10	97.21	96.65	97.49
DeepPermNet (Sinkhorn Norm. + VGG16)	96.87	97.99	96.87	**99.79**	**99.82**	**99.55**	**98.48**
DeepPermNet (Bi-level Opt. + VGG16)	**98.12**	**99.92**	98.13	97.78	98.72	97.87	98.42


Weak
Strong
Bushy-Eyebrows
(Public Figures)
Smiling
(Public Figures)
Open (OSR)
Natural (OSR)

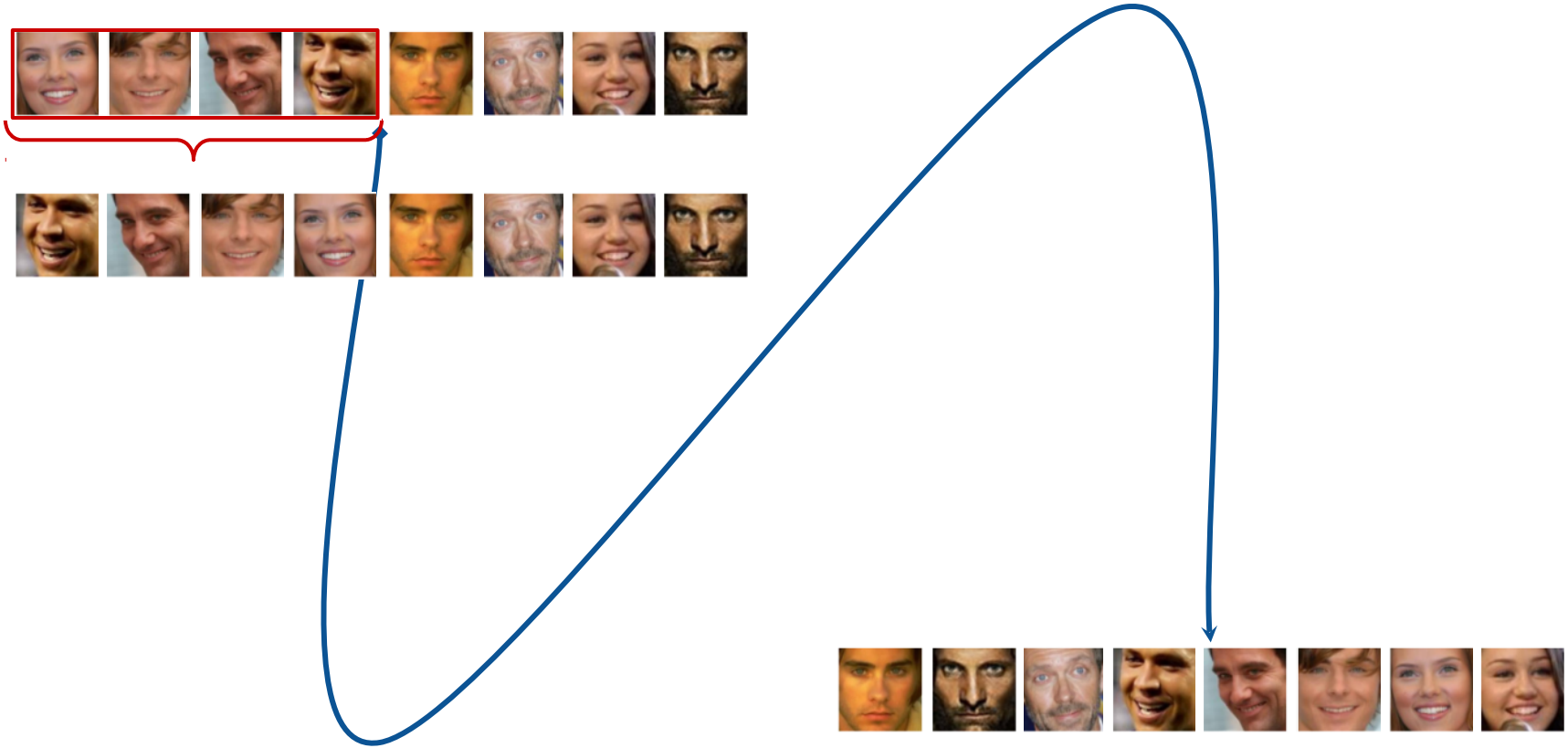
Sorting Long Sequences - Smiling



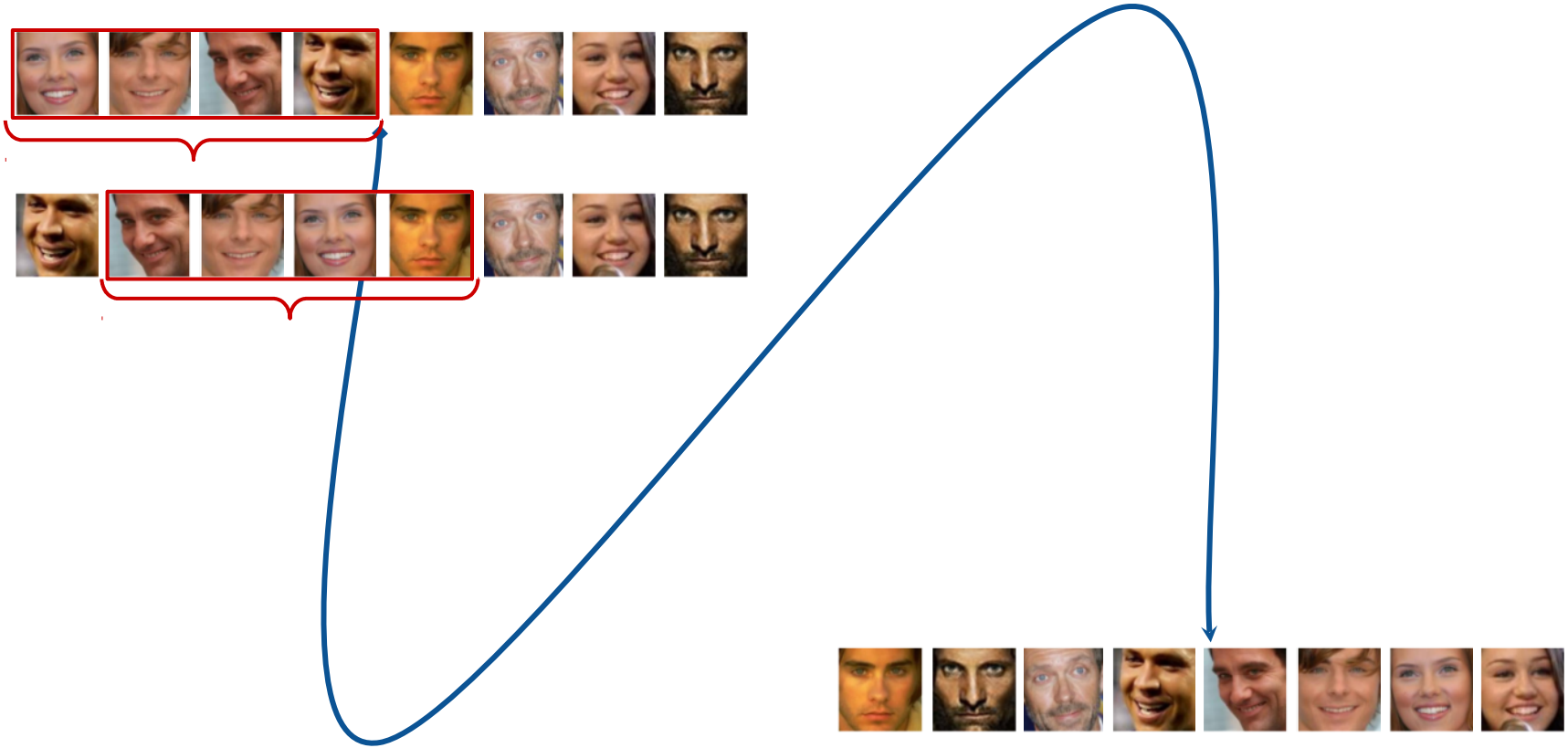
Sorting Long Sequences - Smiling



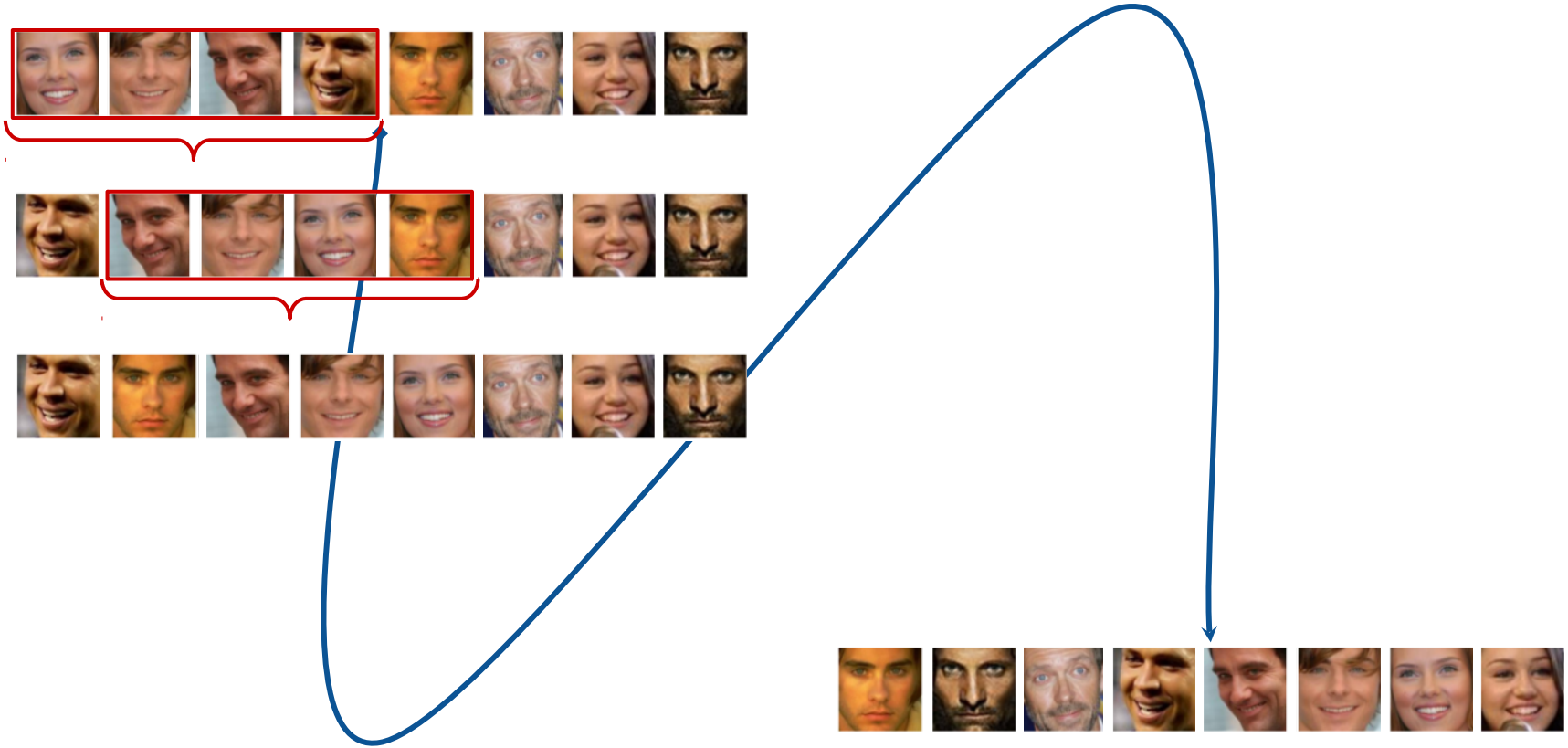
Sorting Long Sequences - Smiling



Sorting Long Sequences - Smiling



Sorting Long Sequences - Smiling



Sorting Long Sequences - Smiling

Sorting Long Sequences - Smiling

Sorting Long Sequences - Smiling

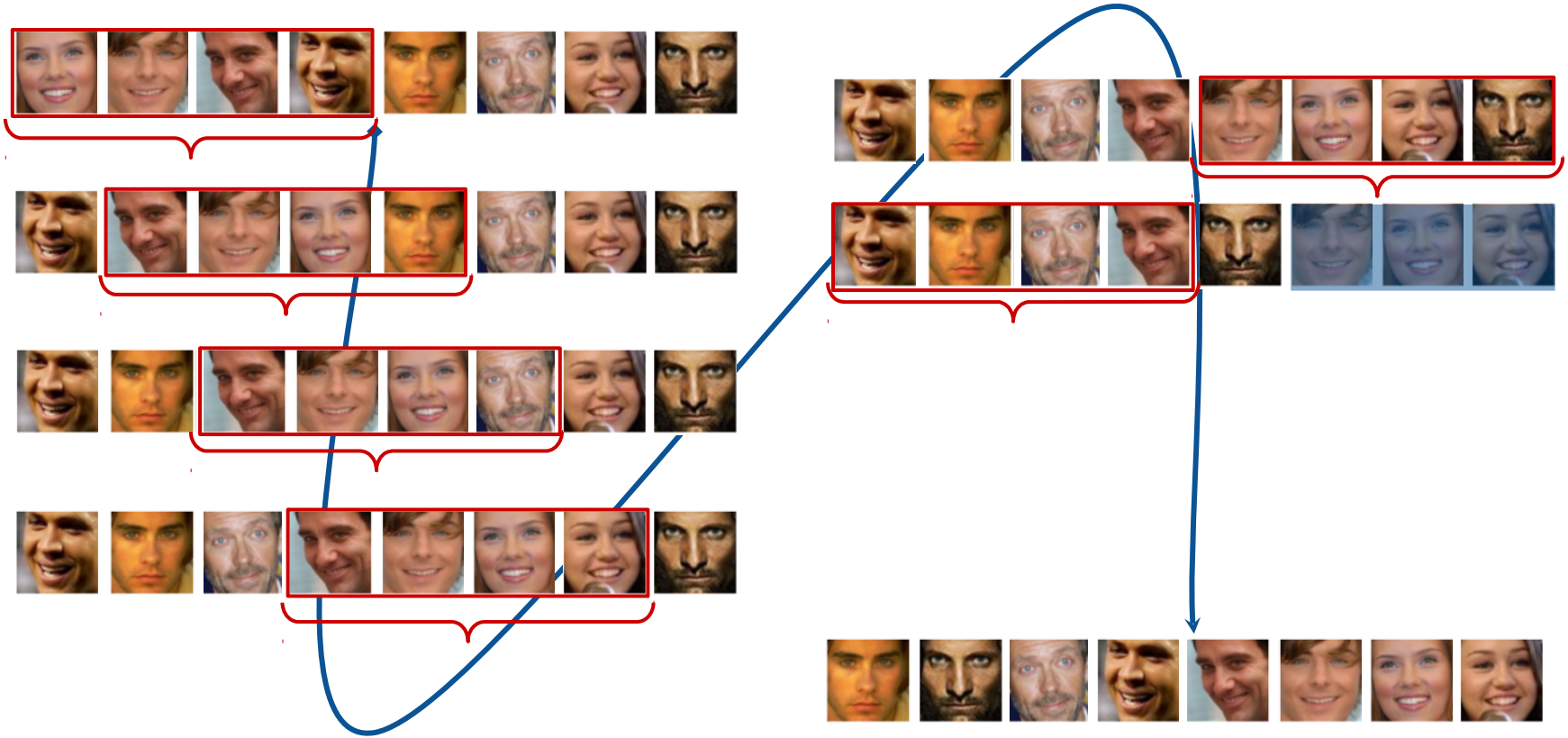
Sorting Long Sequences - Smiling

Sorting Long Sequences - Smiling

Sorting Long Sequences - Smiling

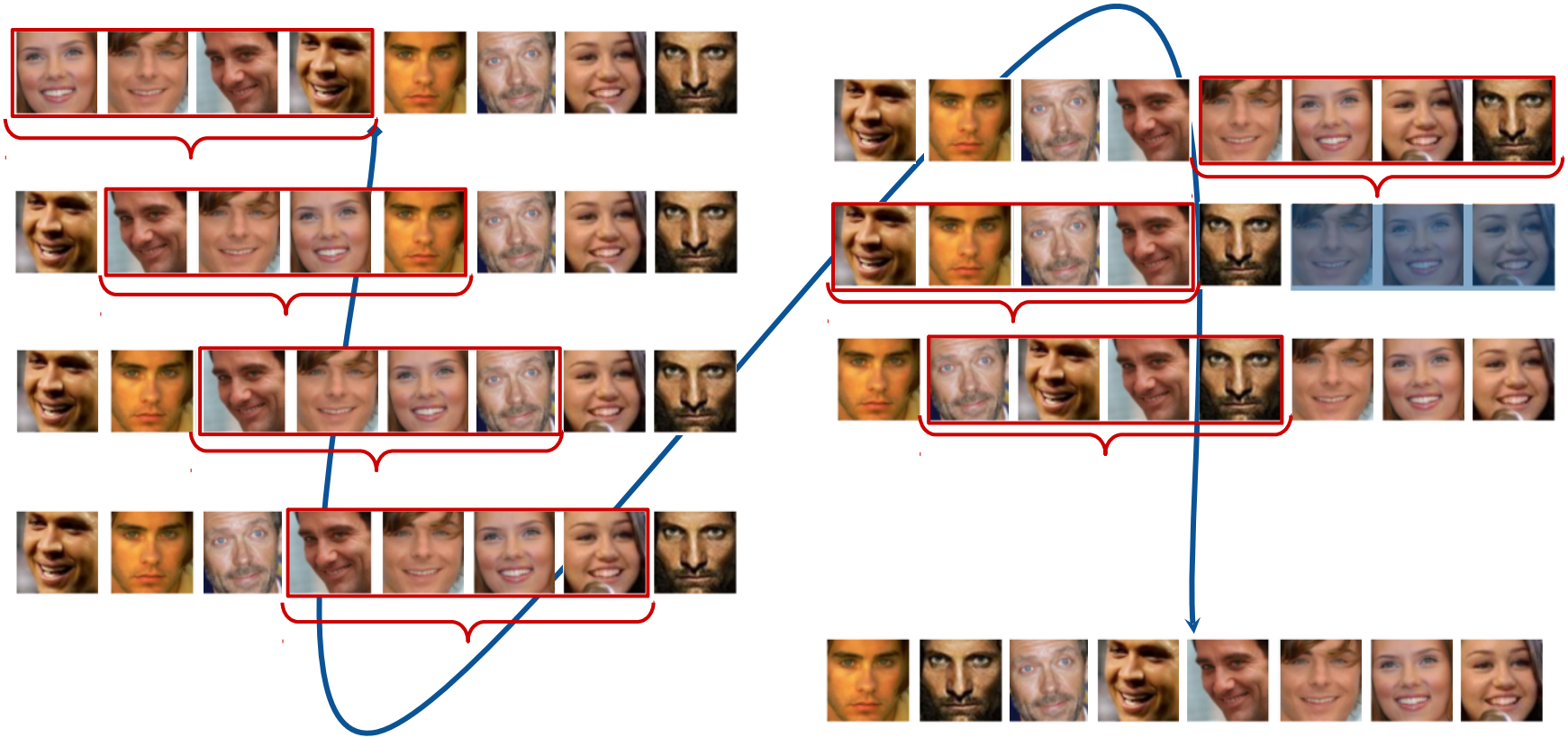
Sorting Long Sequences - Smiling

Sorting Long Sequences - Smiling

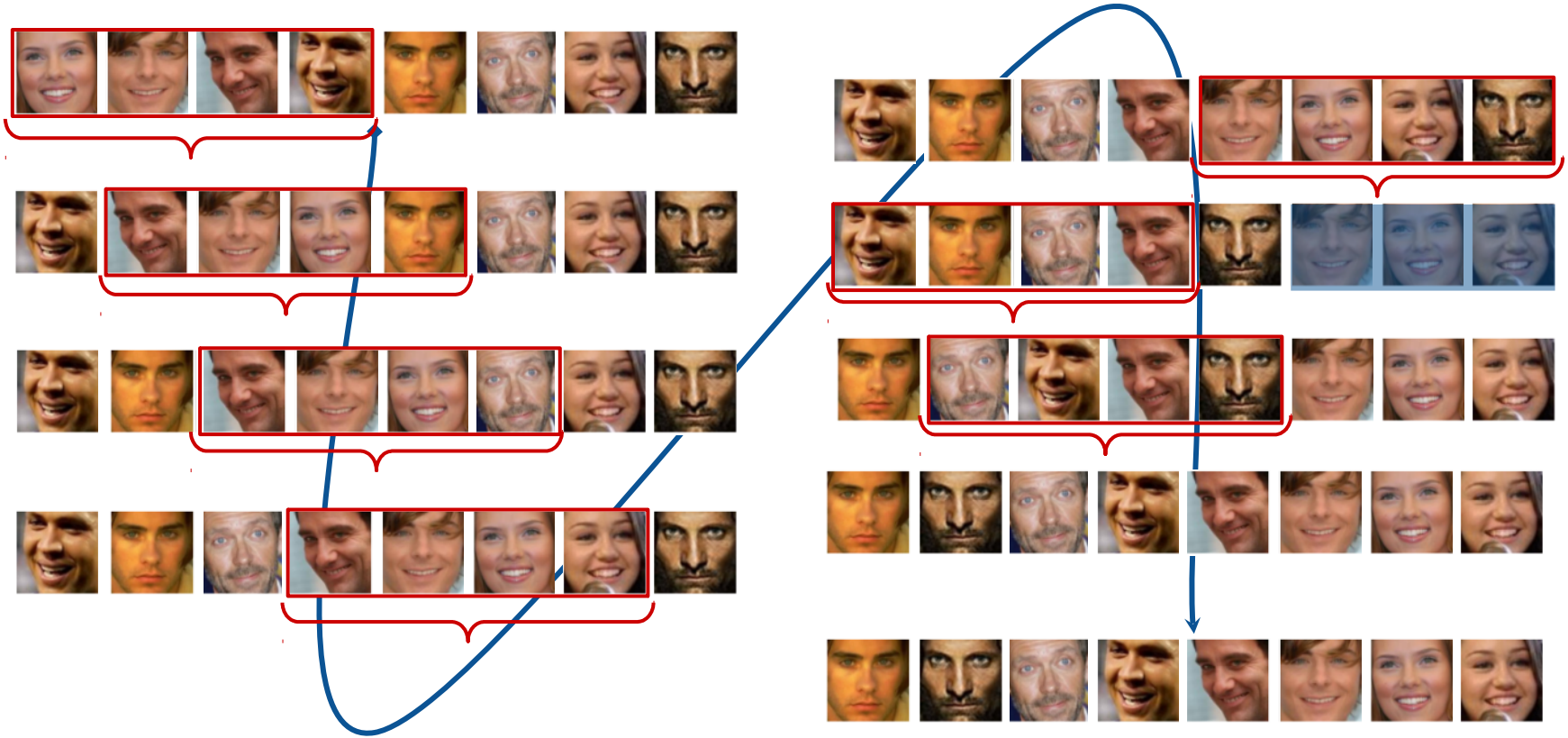


Sorting Long Sequences - Smiling

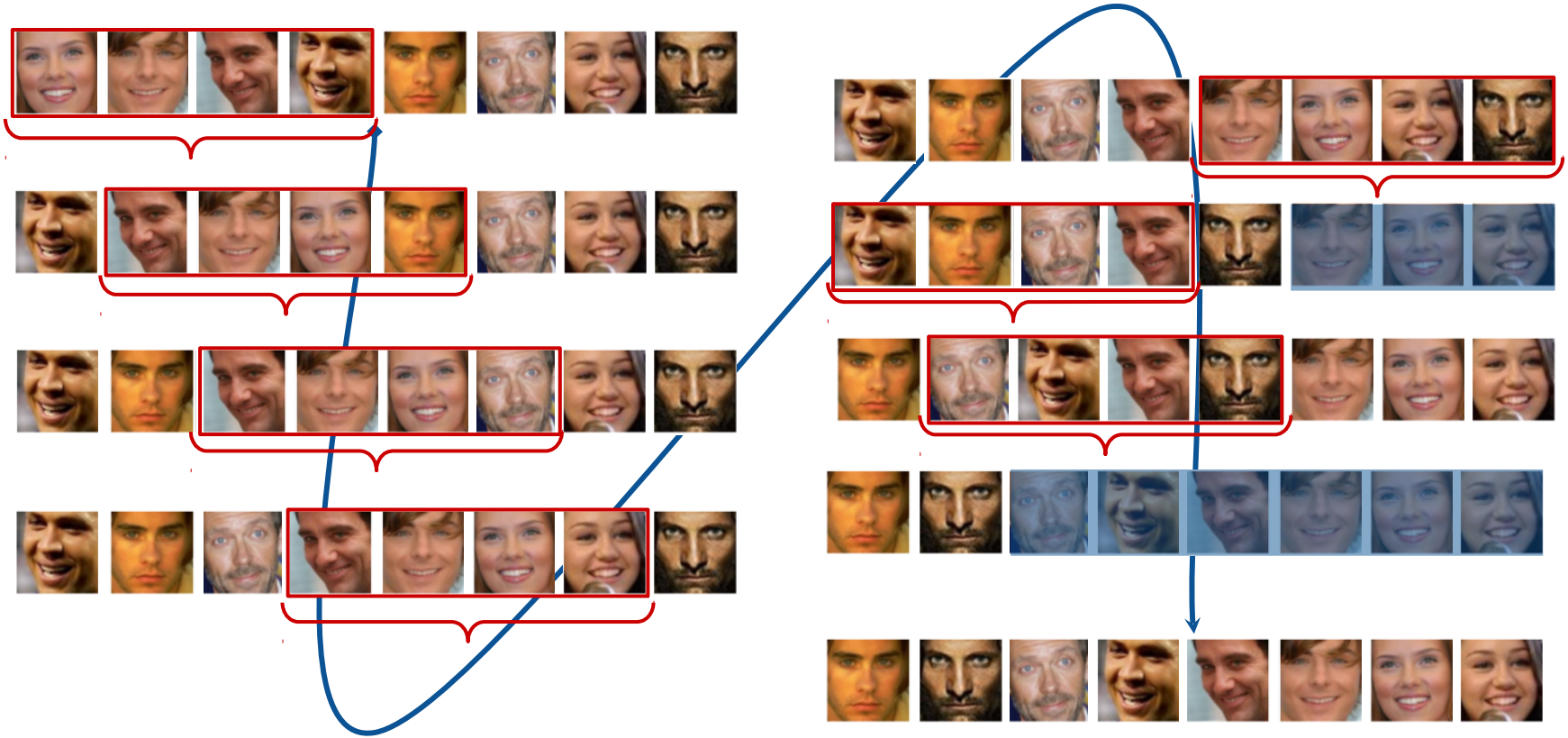
Sorting Long Sequences - Smiling



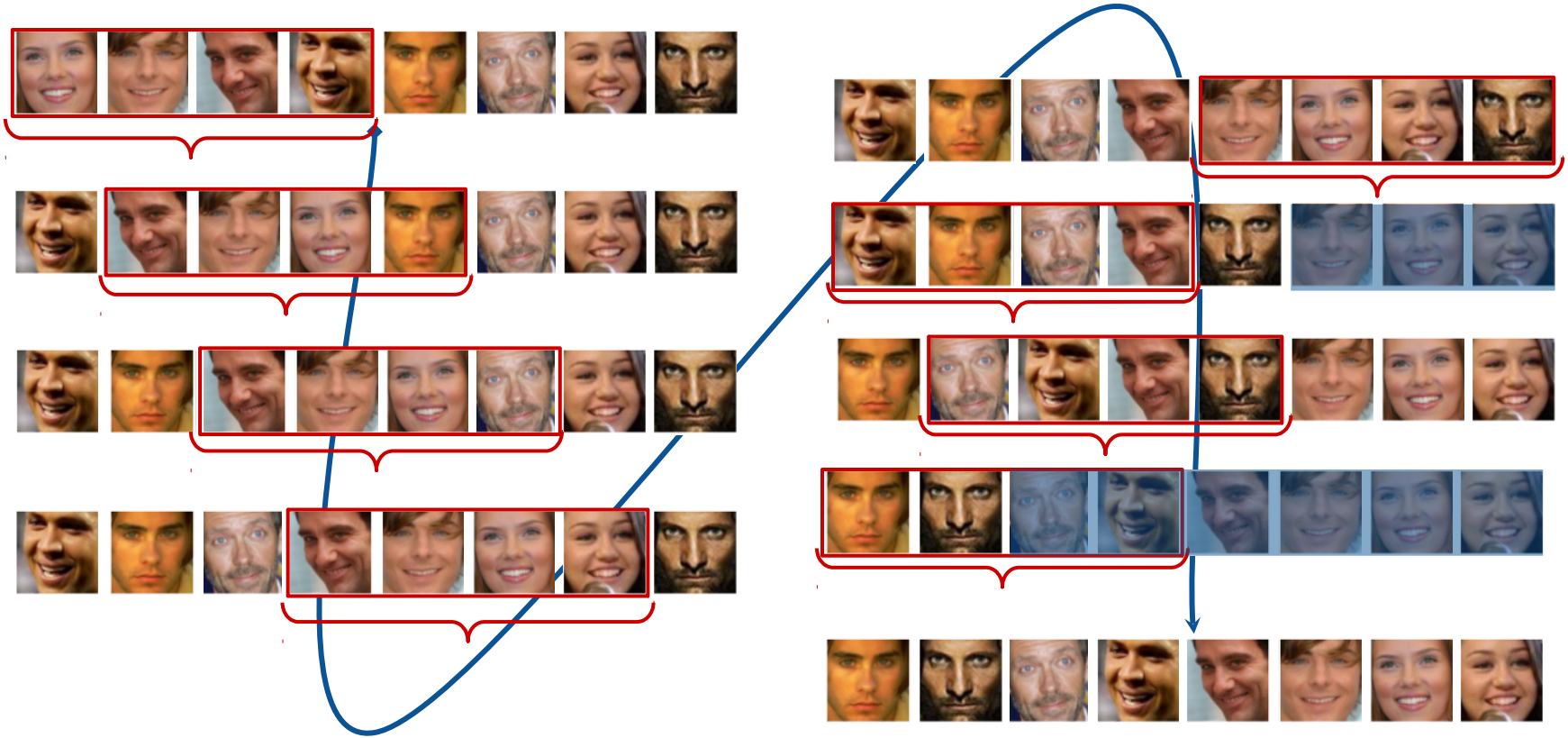
Sorting Long Sequences - Smiling



Sorting Long Sequences - Smiling



Sorting Long Sequences - Smiling



Experiments - Learning to Rank

Permutation prediction + Sorting Algorithm

TABLE 3
Evaluation on supervised learning to rank

Method	Scene interestingness			Car chronology		
	NDCG	KT	Pair. Acc.	NDCG	KT	Pair. Acc.
Joachims [29]	0.870	0.317	65.8	0.928	0.482	74.1
Xu and Li [64]	0.745	-0.077	46.1	0.827	0.118	55.9
Wu et al. [62]	0.860	0.315	64.3	0.935	0.409	70.6
Cao et al. [8]	0.821	0.118	55.9	0.872	0.291	64.5
Xia et al. [63]	0.862	0.282	64.1	0.854	0.278	63.9
Fernando et al. [19]	0.887	0.347	67.4	.949	0.553	76.9
DeepPermNet (Sinkhorn Norm.)	0.922	0.360	68.0	**0.968**	**0.724**	**86.2**
DeepPermNet (Bi-level Opt.)	**0.923**	**0.363**	**68.2**	0.964	0.700	84.9

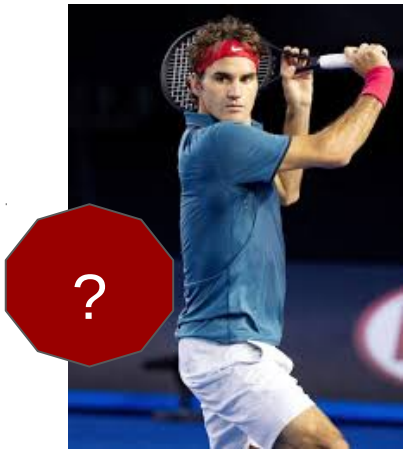
Structure In The Inputs

[DeepPermNet: Visual Permutation Learning. **Rodrigo Santa Cruz**, Basura Fernando, Anoop Cherian, Stephen Gould. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.]

[Visual Permutation Learning. **Rodrigo Santa Cruz**, Basura Fernando, Anoop Cherian, Stephen Gould. Pattern Analysis and Machine Intelligence (PAMI), 2018]

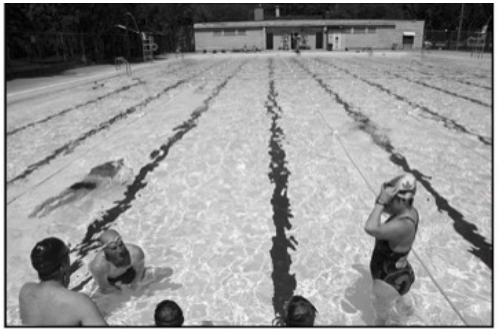
Visual Inputs are structured ...

Visual Inputs are structured ...



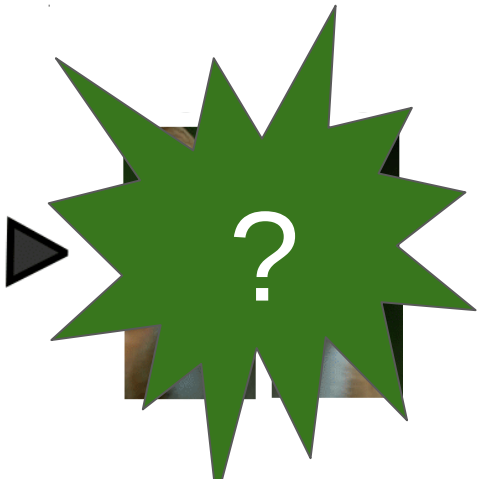
Visual Inputs are structured ...

Visual Inputs are structured ...

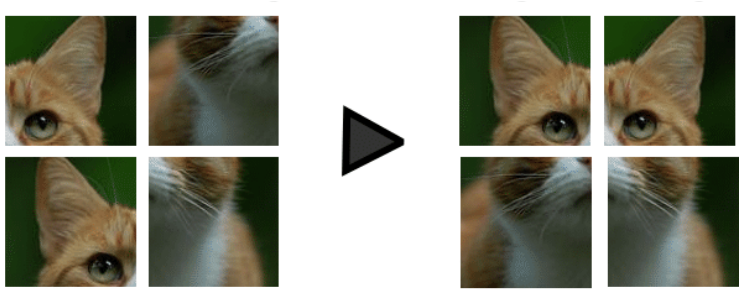


Visual Inputs are structured ...

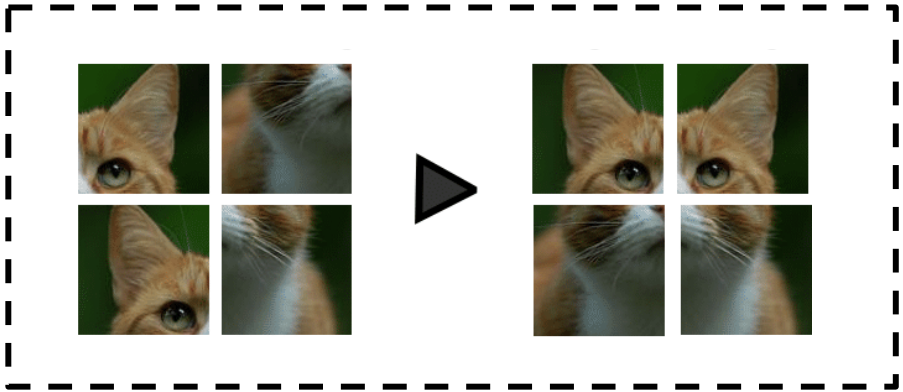
Visual Inputs are structured ...



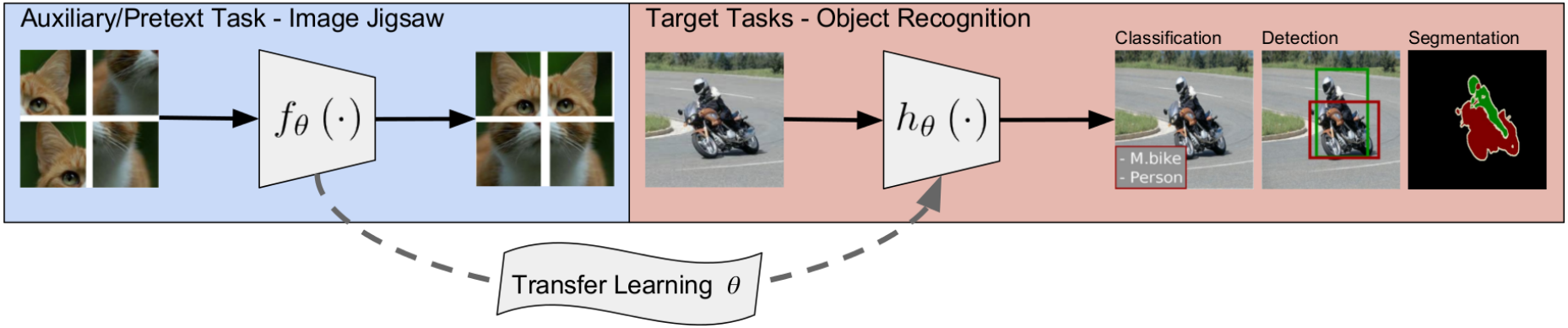
Visual Inputs are structured ...



Visual Inputs are structured ...



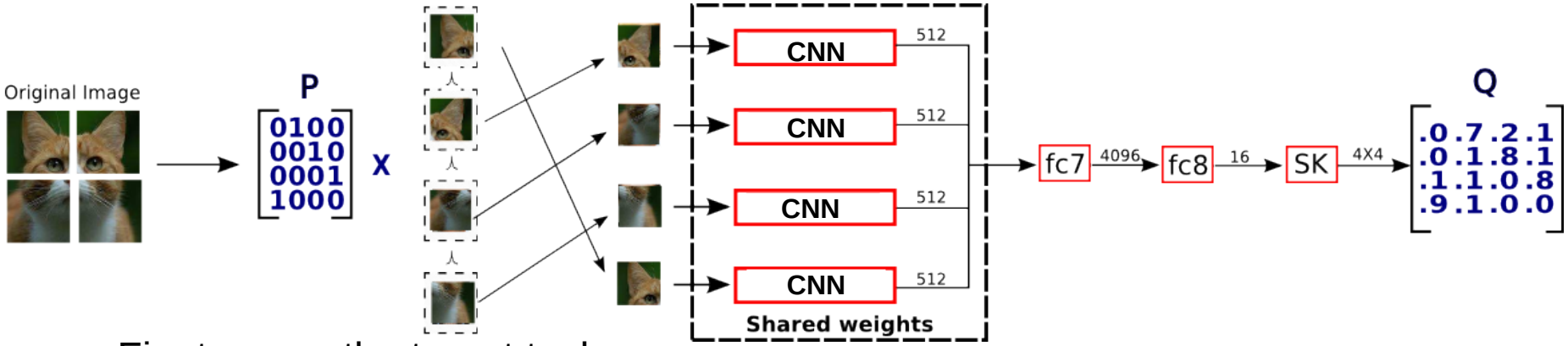
Self-Supervised Representation Learning



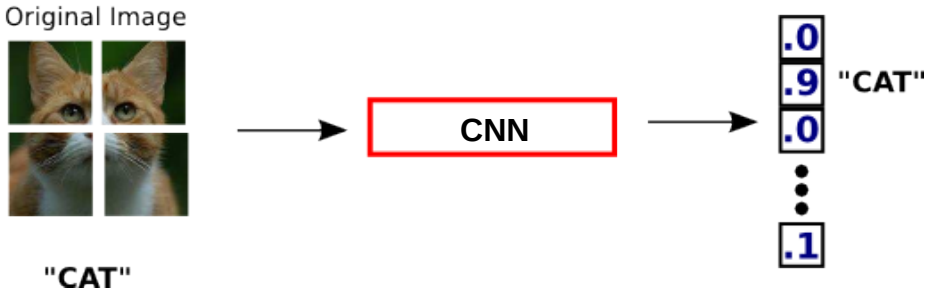
We hypothesize that the model trained to solve such task is able to capture high-level semantic concepts, structure and shared patterns in visual data.

Visual Permutation Learning

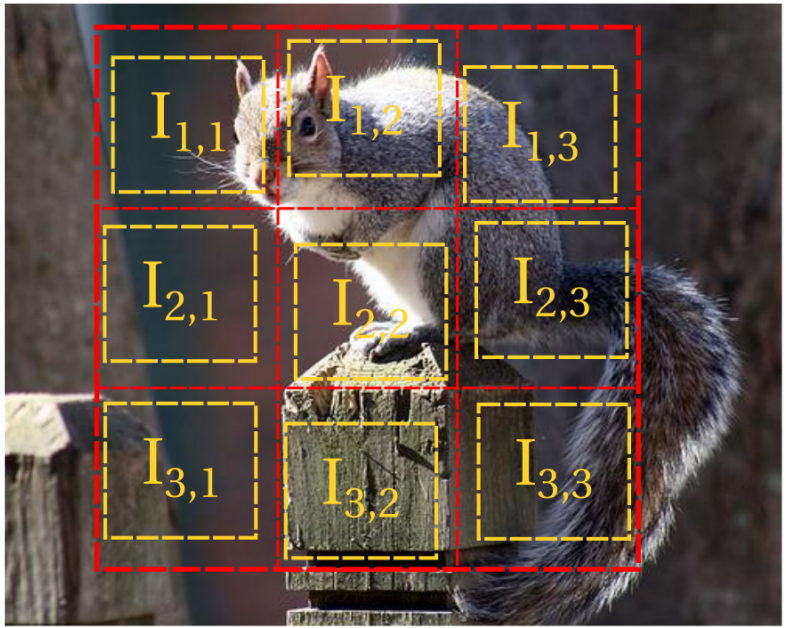
- Pretrain in the visual permutation learning:



- Finetune on the target task:



Avoiding “Shortcuts”



1. Randomly crop a squared region of the image;
2. Split the resized crop into a 3 × 3 grid cell;
3. Randomly select 64 × 64 pixels tiles from each cell;

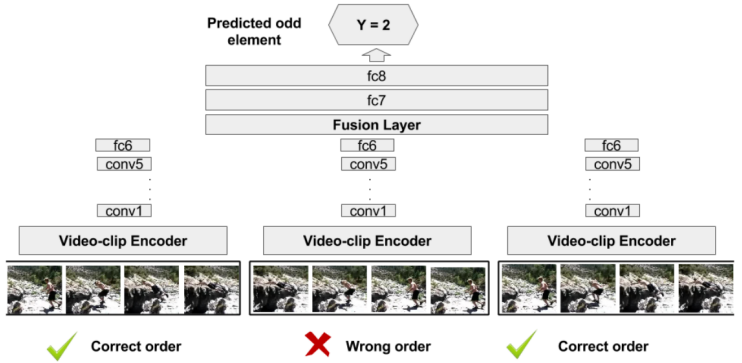
Others: Low level statistics and Chromatic Aberration.

Existing Self-Supervised Learning Methods

[Doersch et al., *ICCV 2015*]

[Zhang et al., *ECCV16*]

[Gidaris, et al., *ICLR 2018*]



[Fernando et al., *ECCV16*]

Pre-training Method	Pretext task	Cls.	Det.	Seg.
ImageNet	Supervised	78.2	56.8	48.0
Random Gaussian	None	53.3	43.4	19.8
Agrawal et al. [2015]	Estimating Ego-motion	52.9	41.8	-
Doersch et al. [2015b]	Context Prediction	55.3	46.6	-
Wang and Gupta [2015]	Visual tracking	58.4	44.0	-
Pathak et al. [2016]	Context autoencoder	56.5	44.5	29.7
Donahue et al. [2017]	Adversarial Learning	58.9	45.7	34.9
Zhang et al. [2016]	Image colorization	65.6	47.9	35.6
Noroozi and Favaro [2016]*	Image jigsaws	67.6	53.2	37.6
Owens et al. [2016]	Ambient sounds	61.3	44.0	-
Bojanowski and Joulin [2017]	Alignment with noisy targets	65.3	49.4	-
Noroozi et al. [2017]	Counting visual primitives	67.7	51.4	36.6
Lee et al. [2017]	Sorting sequences	63.8	46.9	-
Pathak et al. [2017]	Motion-based segmentation	61.0	52.2	-
Zhang et al. [2017b]	Cross-channel prediction	67.1	46.7	36.0
Larsson et al. [2017]	Image colorization	65.9	-	38.0
Jenni and Favaro [2018]	Predicting synthetic artifacts	69.8	52.5	38.1
Gidaris et al. [2018]	Predicting image rotation	**72.97**	54.4	39.1
Kim et al. [2018]	Damaged image jigsaws	69.2	52.4	39.3
Nathan Mundhenk et al. [2018]	Improved context prediction	69.6	**55.8**	**41.2**
Ren and Jae Lee [2018]	Multi-task	68.0	52.6	-
DeepPermNet (Sinkhorn Norm.)*	Visual Permutation Learning	69.4	49.5	37.9
DeepPermNet (Bi-level Opt.)*	Visual Permutation Learning	65.5	45.7	36.4

Pre-training Method	Pretext task	Cls.	Det.	Seg.
ImageNet	Supervised	78.2	56.8	48.0
Random Gaussian	None	53.3	43.4	19.8
Agrawal et al. [2015]	Estimating Ego-motion	52.9	41.8	-
Doersch et al. [2015b]	Context Prediction	55.3	46.6	-
Wang and Gupta [2015]	Visual tracking	58.4	44.0	-
Pathak et al. [2016]	Context autoencoder	56.5	44.5	29.7
Donahue et al. [2017]	Adversarial Learning	58.9	45.7	34.9
Zhang et al. [2016]	Image colorization	65.6	47.9	35.6
Noroozi and Favaro [2016]*	Image jigsaws	67.6	53.2	37.6
Owens et al. [2016]	Ambient sounds	61.3	44.0	-
Bojanowski and Joulin [2017]	Alignment with noisy targets	65.3	49.4	-
Noroozi et al. [2017]	Counting visual primitives	67.7	51.4	36.6
Lee et al. [2017]	Sorting sequences	63.8	46.9	-
Pathak et al. [2017]	Motion-based segmentation	61.0	52.2	-
Zhang et al. [2017b]	Cross-channel prediction	67.1	46.7	36.0
Larsson et al. [2017]	Image colorization	65.9	-	38.0
Jenni and Favaro [2018]	Predicting synthetic artifacts	69.8	52.5	38.1
Gidaris et al. [2018]	Predicting image rotation	**72.97**	54.4	39.1
Kim et al. [2018]	Damaged image jigsaws	69.2	52.4	39.3
Nathan Mundhenk et al. [2018]	Improved context prediction	69.6	**55.8**	**41.2**
Ren and Jae Lee [2018]	Multi-task	68.0	52.6	-
DeepPermNet (Sinkhorn Norm.)*	Visual Permutation Learning	69.4	49.5	37.9
DeepPermNet (Bi-level Opt.)*	Visual Permutation Learning	65.5	45.7	36.4

Δ(Cls) = 16%

Pre-training Method	Pretext task	Cls.	Det.	Seg.
ImageNet	Supervised	78.2	56.8	48.0
Random Gaussian	None	53.3	43.4	19.8
Agrawal et al. [2015]	Estimating Ego-motion	52.9	41.8	-
Doersch et al. [2015b]	Context Prediction	55.3	46.6	-
Wang and Gupta [2015]	Visual tracking	58.4	44.0	-
Pathak et al. [2016]	Context autoencoder	56.5	44.5	29.7
Donahue et al. [2017]	Adversarial Learning	58.9	45.7	34.9
Zhang et al. [2016]	Image colorization	65.6	47.9	35.6
Noroozi and Favaro [2016]*	Image jigsaws	67.6	53.2	37.6
Owens et al. [2016]	Ambient sounds	61.3	44.0	-
Bojanowski and Joulin [2017]	Alignment with noisy targets	65.3	49.4	-
Noroozi et al. [2017]	Counting visual primitives	67.7	51.4	36.6
Lee et al. [2017]	Sorting sequences	63.8	46.9	-
Pathak et al. [2017]	Motion-based segmentation	61.0	52.2	-
Zhang et al. [2017b]	Cross-channel prediction	67.1	46.7	36.0
Larsson et al. [2017]	Image colorization	65.9	-	38.0
Jenni and Favaro [2018]	Predicting synthetic artifacts	69.8	52.5	38.1
Gidaris et al. [2018]	Predicting image rotation	**72.97**	54.4	39.1
Kim et al. [2018]	Damaged image jigsaws	69.2	52.4	39.3
Nathan Mundhenk et al. [2018]	Improved context prediction	69.6	**55.8**	**41.2**
Ren and Jae Lee [2018]	Multi-task	68.0	52.6	-
DeepPermNet (Sinkhorn Norm.)*	Visual Permutation Learning	69.4	49.5	37.9
DeepPermNet (Bi-level Opt.)*	Visual Permutation Learning	65.5	45.7	36.4

Δ(Cls) = 16%

Δ(Det) = 6%

Pre-training Method	Pretext task	Cls.	Det.	Seg.
ImageNet	Supervised	78.2	56.8	48.0
Random Gaussian	None	53.3	43.4	19.8
Agrawal et al. [2015]	Estimating Ego-motion	52.9	41.8	-
Doersch et al. [2015b]	Context Prediction	55.3	46.6	-
Wang and Gupta [2015]	Visual tracking	58.4	44.0	-
Pathak et al. [2016]	Context autoencoder	56.5	44.5	29.7
Donahue et al. [2017]	Adversarial Learning	58.9	45.7	34.9
Zhang et al. [2016]	Image colorization	65.6	47.9	35.6
Noroozi and Favaro [2016]*	Image jigsaws	67.6	53.2	37.6
Owens et al. [2016]	Ambient sounds	61.3	44.0	-
Bojanowski and Joulin [2017]	Alignment with noisy targets	65.3	49.4	-
Noroozi et al. [2017]	Counting visual primitives	67.7	51.4	36.6
Lee et al. [2017]	Sorting sequences	63.8	46.9	-
Pathak et al. [2017]	Motion-based segmentation	61.0	52.2	-
Zhang et al. [2017b]	Cross-channel prediction	67.1	46.7	36.0
Larsson et al. [2017]	Image colorization	65.9	-	38.0
Jenni and Favaro [2018]	Predicting synthetic artifacts	69.8	52.5	38.1
Gidaris et al. [2018]	Predicting image rotation	**72.97**	54.4	39.1
Kim et al. [2018]	Damaged image jigsaws	69.2	52.4	39.3
Nathan Mundhenk et al. [2018]	Improved context prediction	69.6	**55.8**	**41.2**
Ren and Jae Lee [2018]	Multi-task	68.0	52.6	-
DeepPermNet (Sinkhorn Norm.)*	Visual Permutation Learning	69.4	49.5	37.9
DeepPermNet (Bi-level Opt.)*	Visual Permutation Learning	65.5	45.7	36.4

Δ(Cls) = 16%

Δ(Det) = 6%

Δ(Seg) = 18%

Pre-training Method	Pretext task	Cls.	Det.	Seg.
ImageNet	Supervised	78.2	56.8	48.0
Random Gaussian	None	53.3	43.4	19.8
Agrawal et al. [2015]	Estimating Ego-motion	52.9	41.8	-
Doersch et al. [2015b]	Context Prediction	55.3	46.6	-
Wang and Gupta [2015]	Visual tracking	58.4	44.0	-
Pathak et al. [2016]	Context autoencoder	56.5	44.5	29.7
Donahue et al. [2017]	Adversarial Learning	58.9	45.7	34.9
Zhang et al. [2016]	Image colorization	65.6	47.9	35.6
Noroozi and Favaro [2016]*	Image jigsaws	67.6	53.2	37.6
Owens et al. [2016]	Ambient sounds	61.3	44.0	-
Bojanowski and Joulin [2017]	Alignment with noisy targets	65.3	49.4	-
Noroozi et al. [2017]	Counting visual primitives	67.7	51.4	36.6
Lee et al. [2017]	Sorting sequences	63.8	46.9	-
Pathak et al. [2017]	Motion-based segmentation	61.0	52.2	-
Zhang et al. [2017b]	Cross-channel prediction	67.1	46.7	36.0
Larsson et al. [2017]	Image colorization	65.9	-	38.0
Jenni and Favaro [2018]	Predicting synthetic artifacts	69.8	52.5	38.1
Gidaris et al. [2018]	Predicting image rotation	**72.97**	54.4	39.1
Kim et al. [2018]	Damaged image jigsaws	69.2	52.4	39.3
Nathan Mundhenk et al. [2018]	Improved context prediction	69.6	**55.8**	**41.2**
Ren and Jae Lee [2018]	Multi-task	68.0	52.6	-
DeepPermNet (Sinkhorn Norm.)*	Visual Permutation Learning	69.4	49.5	37.9
DeepPermNet (Bi-level Opt.)*	Visual Permutation Learning	65.5	45.7	36.4

Δ(Cls) = 16%

Δ(Det) = 6%

Δ(Seg) = 18%

Structure In The Models

[Neural Algebra of Classifiers. **Rodrigo Santa Cruz**, Basura Fernando, Anoop Cherian, Stephen Gould. In IEEE Winter Conference on Applications of Computer Vision (WACV), 2018.]

Compositional Model

Compositional Model

Which one is an **albatross**?

Compositional Model

Which one is an **albatross**?
Albatrosses are birds with hooked beak and large wingspan.




Compositional Model

Which one is an **albatross**?
Albatrosses are birds with hooked beak and large wingspan.

Albatross

Compositional Model

Which one is an **albatross**?
Albatrosses are birds with hooked beak and large wingspan.
Which one is a **frigatebird**?

Albatross

Compositional Model

Which one is an **albatross**?
Albatrosses are birds with hooked beak and large wingspan.
Which one is a **frigatebird**?
Frigatebirds seem black albatrosses with white or red pouch.

Albatross

Compositional Model

Which one is an **albatross**?
Albatrosses are birds with hooked beak and large wingspan.
Which one is a **frigatebird**?
Frigatebirds seem black albatrosses with white or red pouch.

Albatross

Frigatebird

Compositional Model

Which one is an **albatross**?
Albatrosses are birds with hooked beak and large wingspan.
Which one is a **frigatebird**?
Frigatebirds seem black albatrosses with white or red pouch.

Albatross

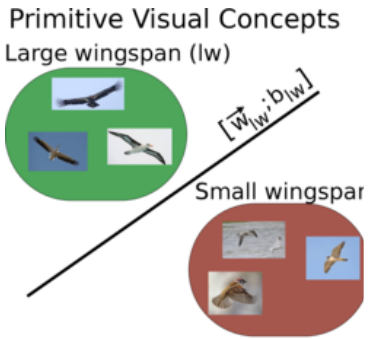
Frigatebird

The **human recognition system** is fundamentally **compositional**, so **unseen visual complex concepts** are recognized from the composition of **simple visual primitives** according to **well-defined rules**.

Neural Algebra Of Classifiers

Neural Algebra Of Classifiers

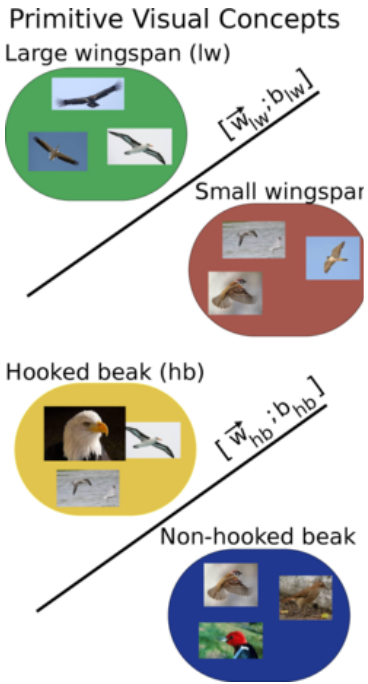
Large wingspan (lw)



Neural Algebra Of Classifiers

Large wingspan (lw)

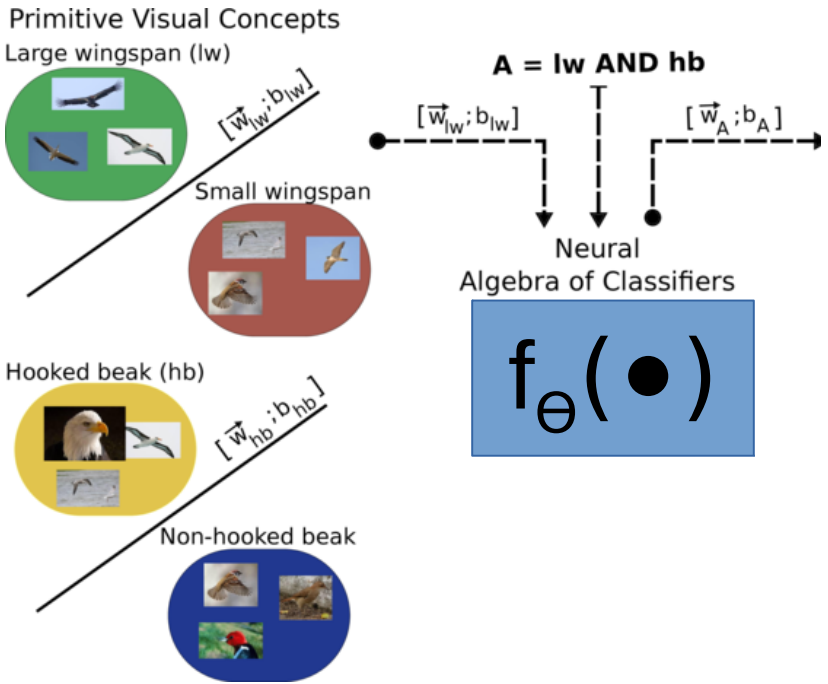
Hooked beak (hb)



Neural Algebra Of Classifiers

Large wingspan (lw)

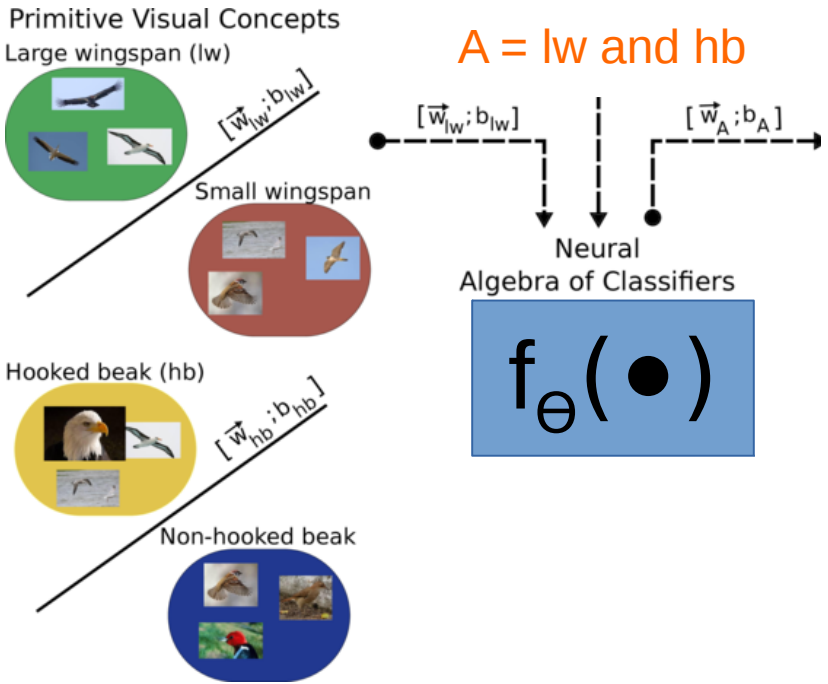
Hooked beak (hb)



Neural Algebra Of Classifiers

Large wingspan (lw)

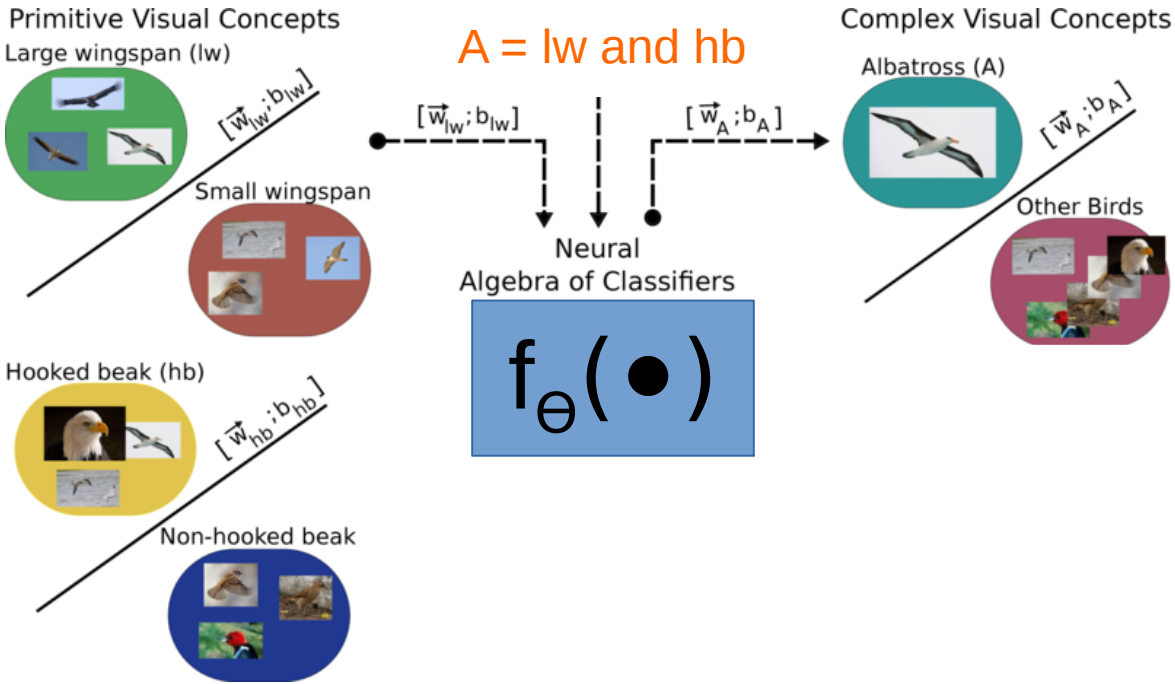
Hooked beak (hb)



Neural Algebra Of Classifiers

Large wingspan (lw)

Hooked beak (hb)

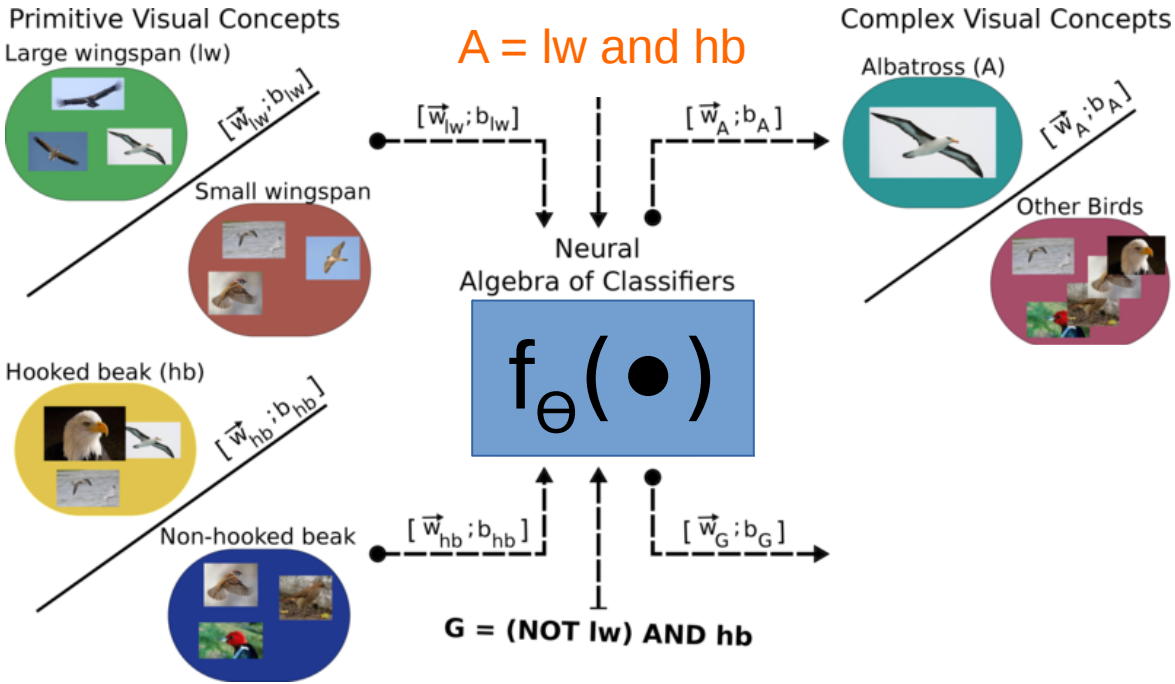


Albatross (A)

Neural Algebra Of Classifiers

Large wingspan (lw)

Hooked beak (hb)

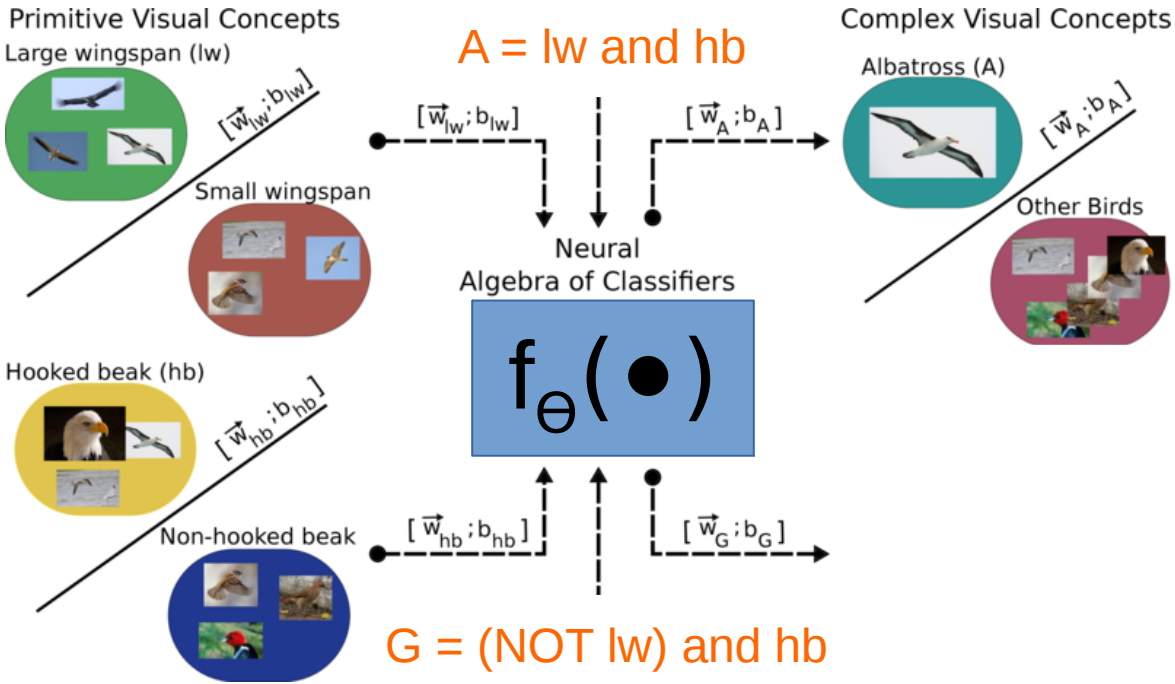


Albatross (A)

Neural Algebra Of Classifiers

Large wingspan (lw)

Hooked beak (hb)

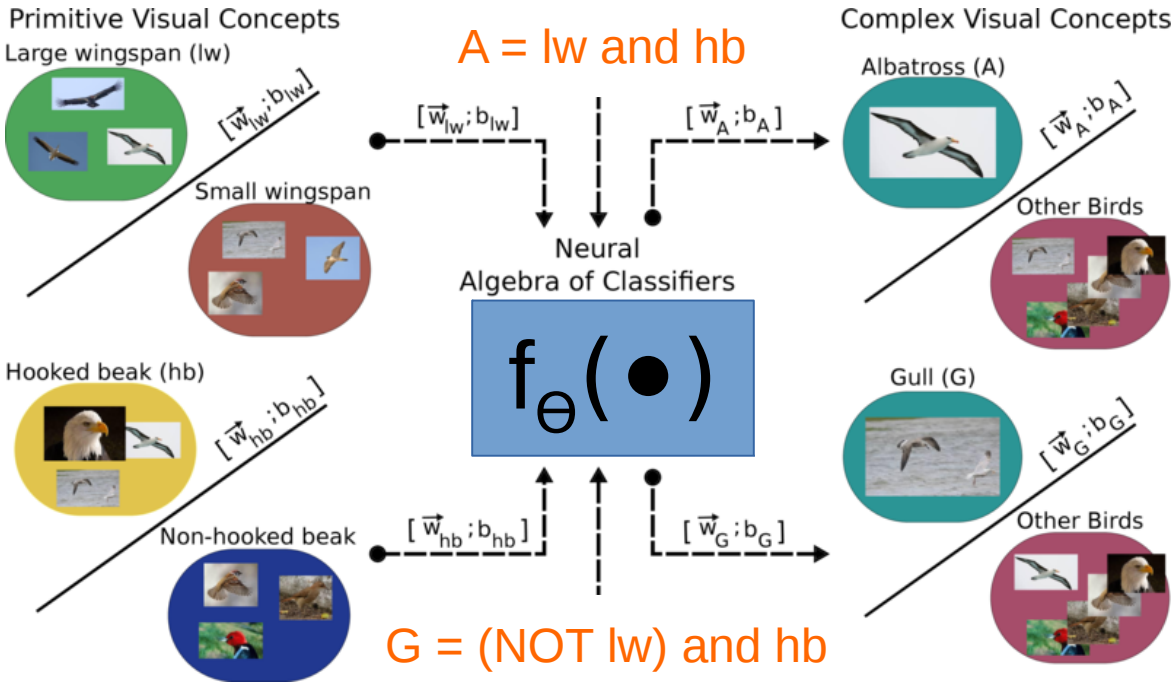


Albatross (A)

Neural Algebra Of Classifiers

Large wingspan (lw)

Hooked beak (hb)



Albatross (A)

Gull (G)

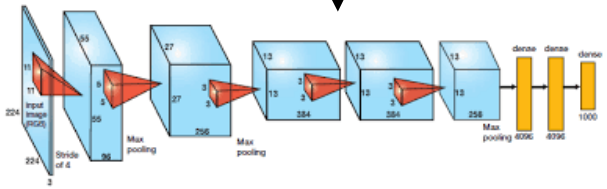
Closed World Assumption

Training Data

person



Test Data



Person

Closed World Assumption

Training Data

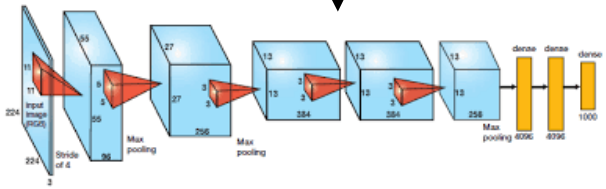
person



car



Test Data



Person
Car

Closed World Assumption

Training Data

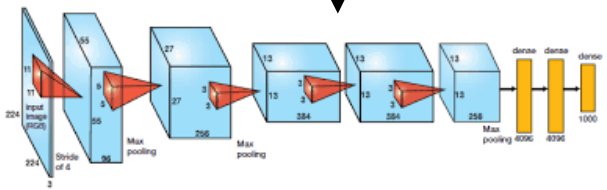
person

car

cat

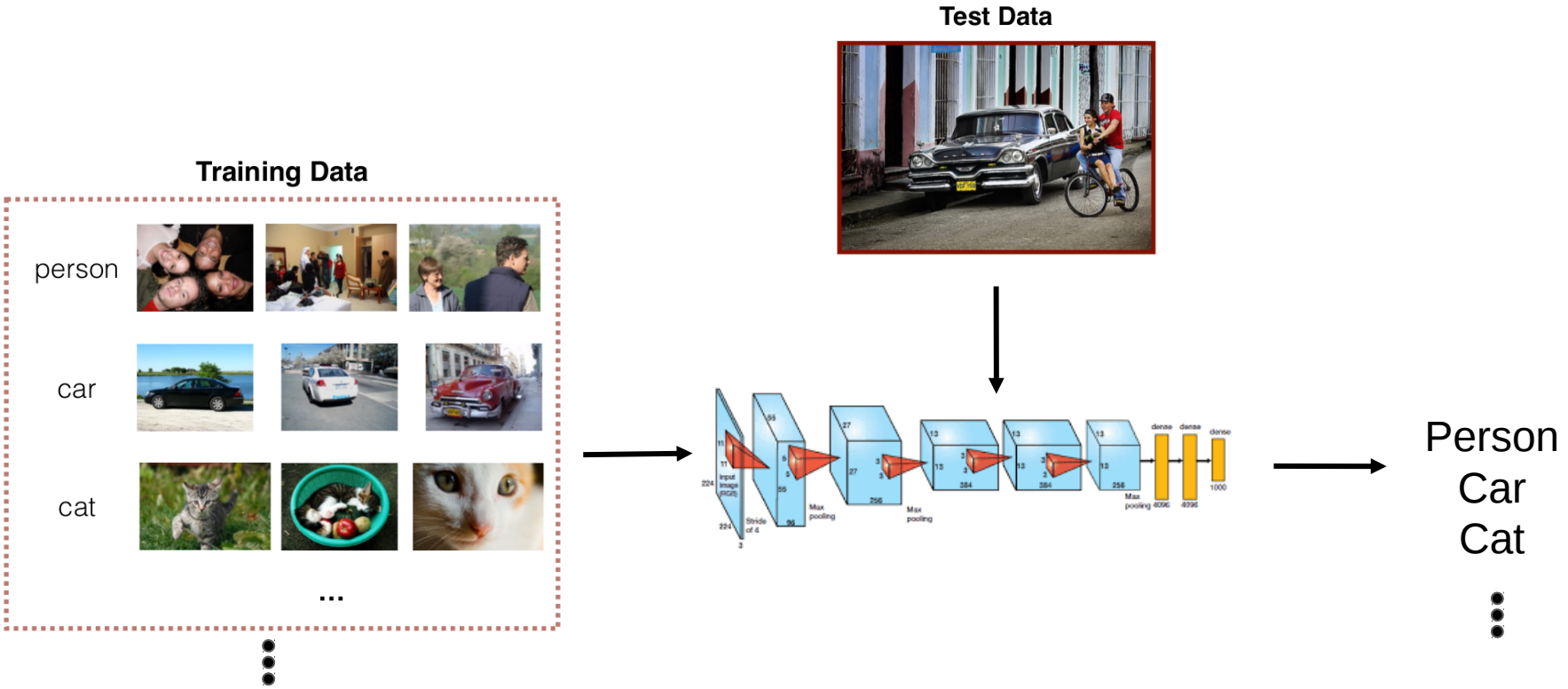
...

Test Data



Person
Car
Cat

Closed World Assumption



Learning

We propose to learn a function $f_\Theta(\cdot)$ that maps the space of expressions to the space of binary classifiers:

Learning

We propose to learn a function $f_\Theta(\cdot)$ that maps the space of expressions to the space of binary classifiers:

Training:

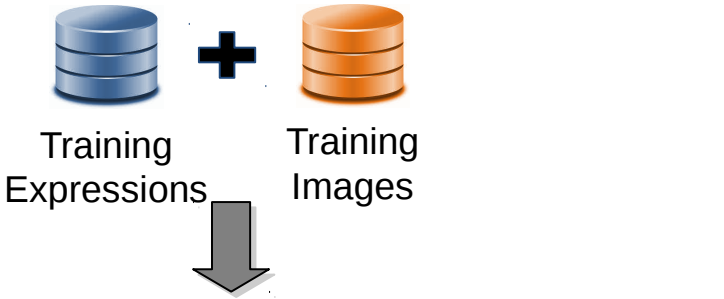
Training Expressions

Training Images

Learning

We propose to learn a function $f_\Theta(\cdot)$ that maps the space of expressions to the space of binary classifiers:

Training:



Learn...

1) Primitives: $w_p \in R^D$

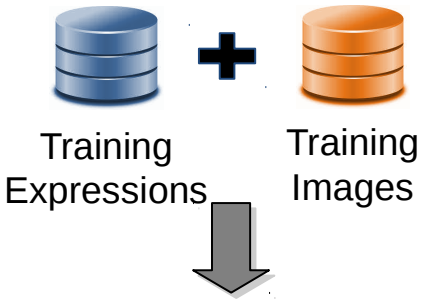
2) Image representation: $h_\Phi(x) \in R^D$

3) Mapping function: $f_\Theta(e) \in R^D$

Learning

We propose to learn a function $f_\Theta(\cdot)$ that maps the space of expressions to the space of binary classifiers:

Training:



Learn...

1) Primitives: $w_p \in R^D$
2) Image representation: $h_\Phi(x) \in R^D$
3) Mapping function: $f_\Theta(e) \in R^D$

Test:

Learning

We propose to learn a function $f_\Theta(\cdot)$ that maps the space of expressions to the space of binary classifiers:

Training:

Training Expressions

Training Images

Learn...

1) Primitives: $w_p \in R^D$
2) Image representation: $h_\Phi(x) \in R^D$
3) Mapping function: $f_\Theta(e) \in R^D$

Test:

Training Expressions

Validation Images

Learning

We propose to learn a function $f_\Theta(\cdot)$ that maps the space of expressions to the space of binary classifiers:

Training:

Training Expressions

Training Images

Learn...

1) Primitives: $w_p \in R^D$
2) Image representation: $h_\Phi(x) \in R^D$
3) Mapping function: $f_\Theta(e) \in R^D$

Test:

Training Expressions

Validation Images

Test Expressions

Test Images

Learning

We propose to learn a function $f_\Theta(\cdot)$ that maps the space of expressions to the space of binary classifiers:

Training:

Training Expressions

Training Images

Learn...

1) Primitives: $w_p \in R^D$
2) Image representation: $h_\Phi(x) \in R^D$
3) Mapping function: $f_\Theta(e) \in R^D$

Test:

Training Expressions

Validation Images

Test Expressions

Test Images

Compute...

1) Image representation $h_\Phi(x)$
2) Classifier $f_\Theta(e)$
3) Score $f_\Theta(e)^T h_\Phi(x)$

Learning

We propose to learn a function $f_\Theta(\cdot)$ that maps the space of expressions to the space of binary classifiers:

Training:

Training Expressions

Training Images

Learn...

1) Primitives: $w_p \in R^D$
2) Image representation: $h_\Phi(x) \in R^D$
3) Mapping function: $f_\Theta(e) \in R^D$

Test:

Training Expressions

Validation Images

Test Expressions

Test Images

Compute...

1) Image representation $h_\Phi(x)$
2) Classifier $f_\Theta(e)$
3) Score $f_\Theta(e)^T h_\Phi(x)$

Learning

We propose to learn a function $f_\Theta(\cdot)$ that maps the space of expressions to the space of binary classifiers:

Training:

Training Expressions

Training Images

Learn...

1) Primitives: $w_p \in R^D$
2) Image representation: $h_\Phi(x) \in R^D$
3) Mapping function: $f_\Theta(e) \in R^D$

Generalize

Test:

Training Expressions

Validation Images

Test Expressions

Test Images

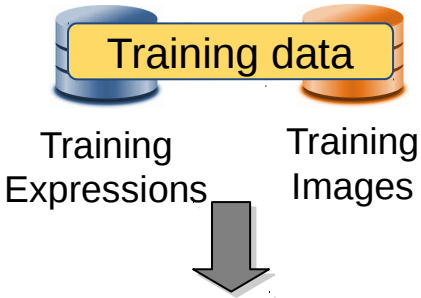
Compute...

1) Image representation $h_\Phi(x)$
2) Classifier $f_\Theta(e)$
3) Score $f_\Theta(e)^T h_\Phi(x)$

Learning

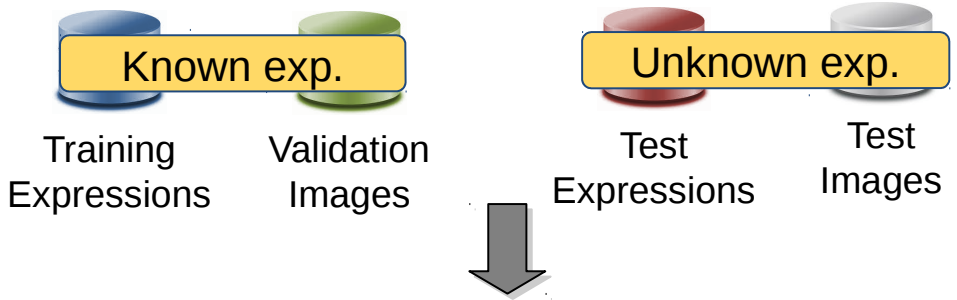
We propose to learn a function $f_\Theta(\cdot)$ that maps the space of expressions to the space of binary classifiers:

Training:

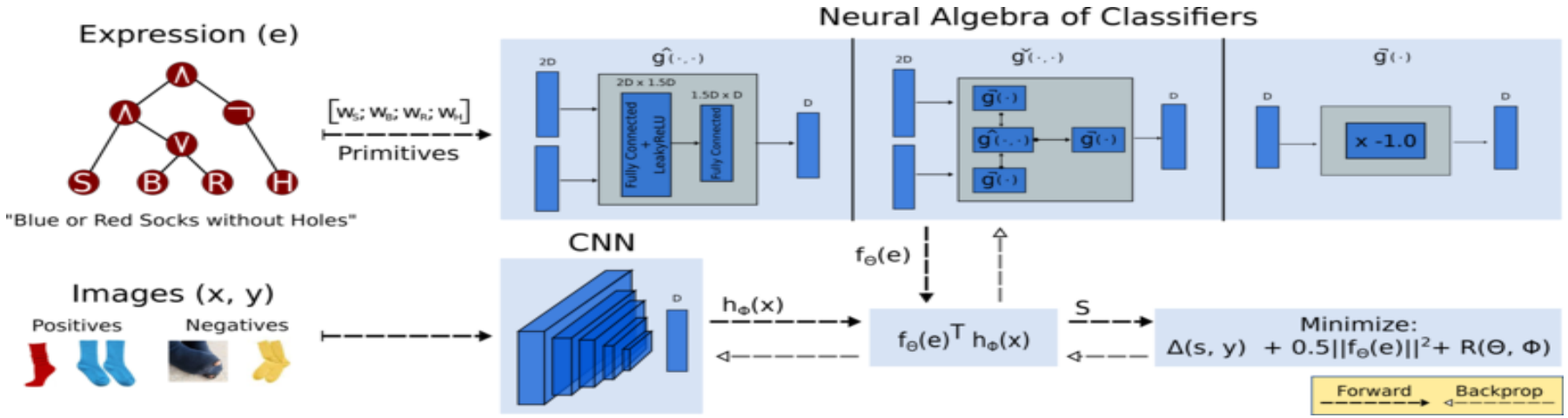


Generalize

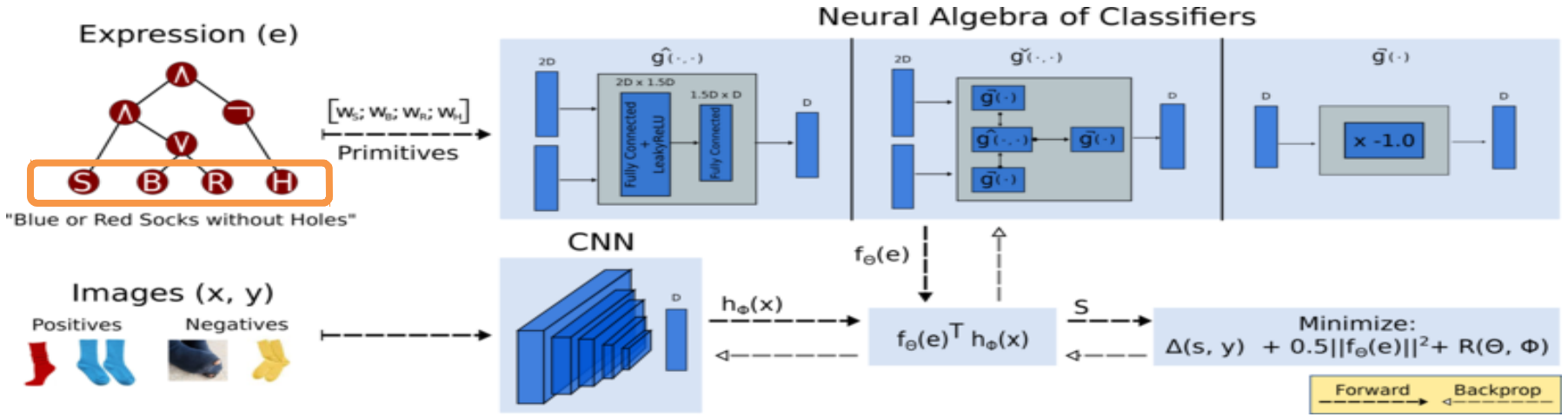
Test:



Model

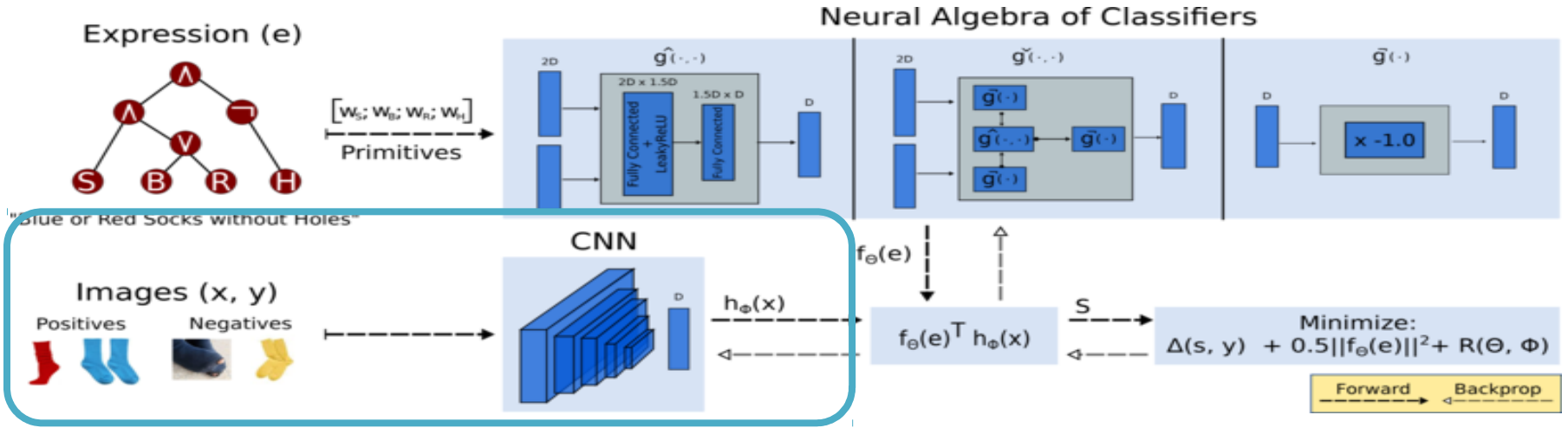

Expression (e)
"Blue or Red Socks without Holes"
Primitives
Neural Algebra of Classifiers
CNN
Images (x, y)
Positives
Negatives
Fully Connected + LeakyReLU
Fully Connected
x -1.0
Minimize:
Δ(s, y) + 0.5||f_Θ(e)||²+ R(Θ, Φ)
Forward
Backprop

Model



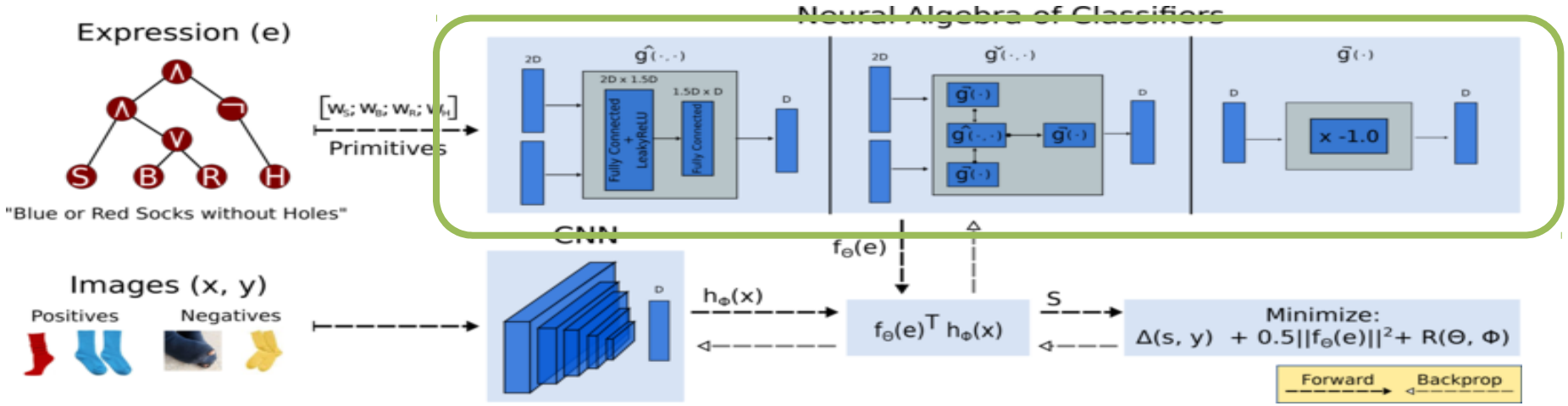
We represent primitives by the parameters of one-vs-all SVM classifiers trained on positives and negatives images of the primitives.

Model



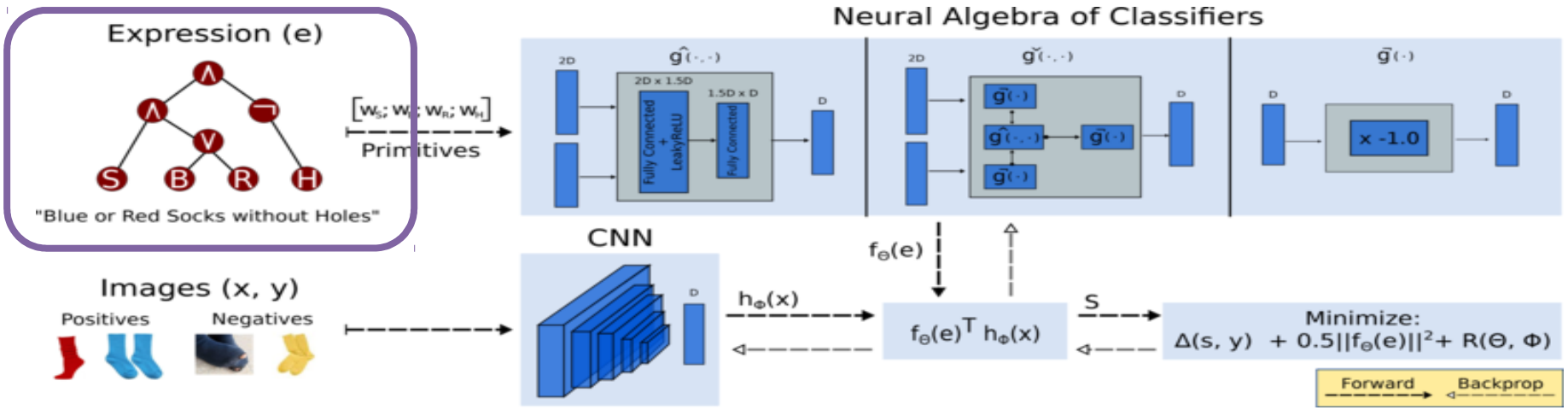
We represent images in a feature space, e.g., CNN features.

Model



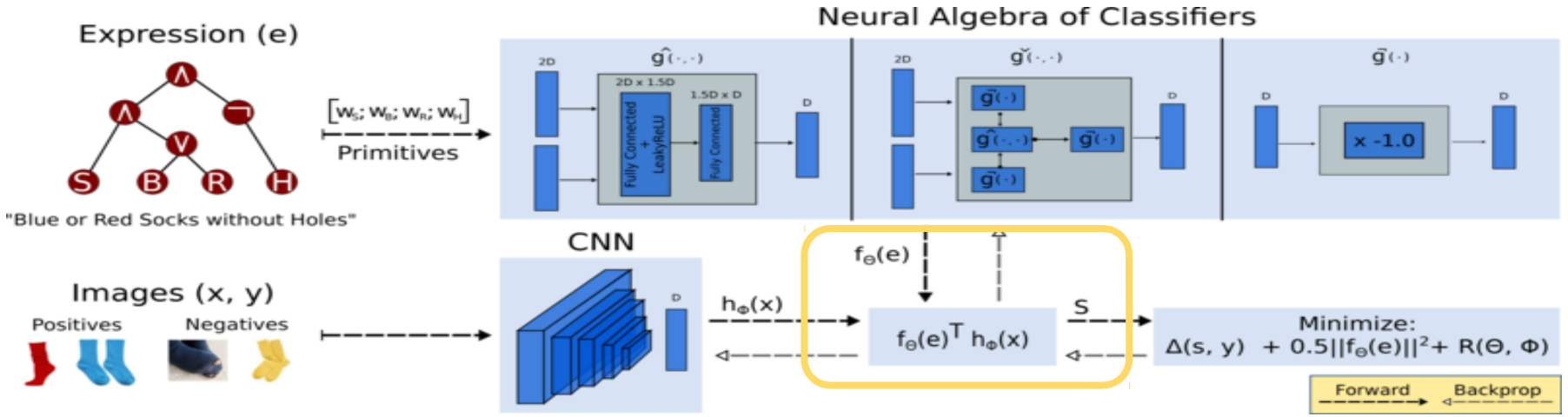
We model our function as a set of composition functions and simplify them using simple analytical relations and De Morgan's laws.

Model



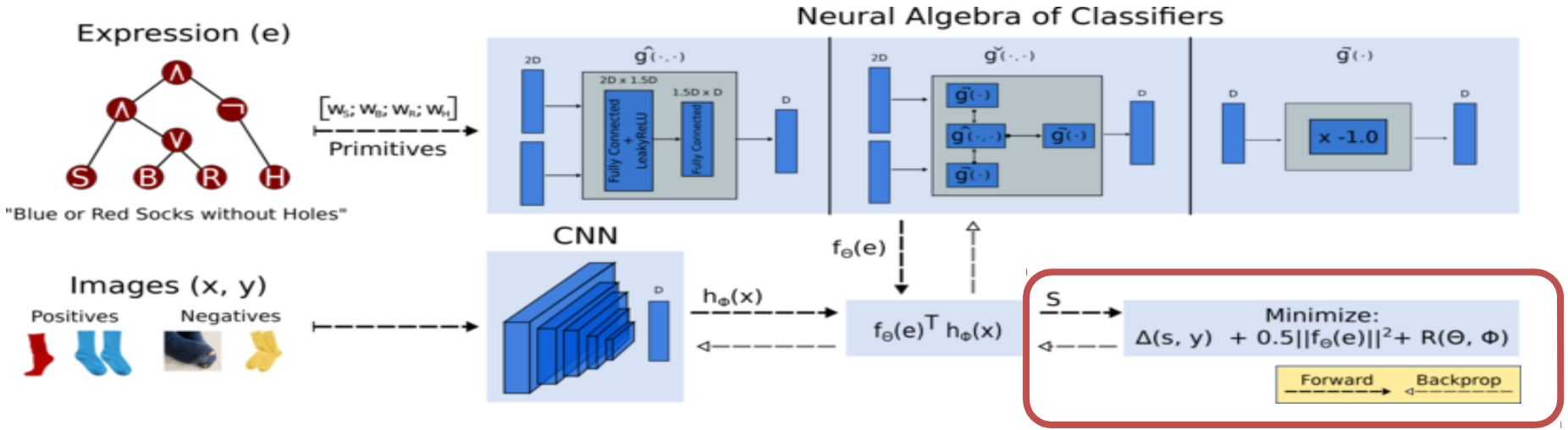
We parse the expression tree applying the composition functions recursively.

Model



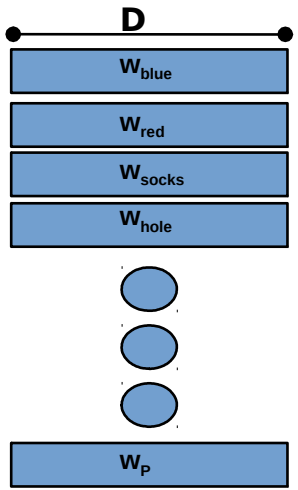
We score images according to the “predicted classifier” for a given expressions.

Model

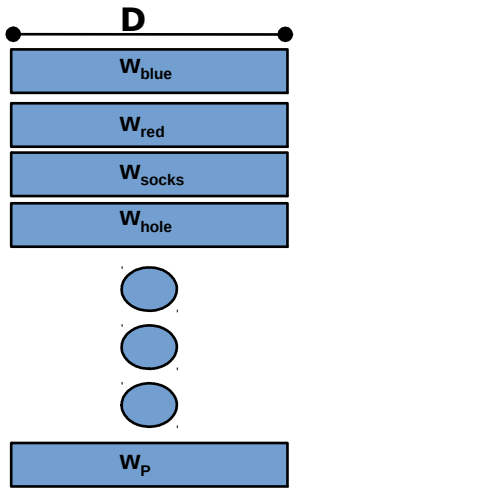


We minimize the classification loss of batches of positive and negative images for different training expressions.

Inference



Inference

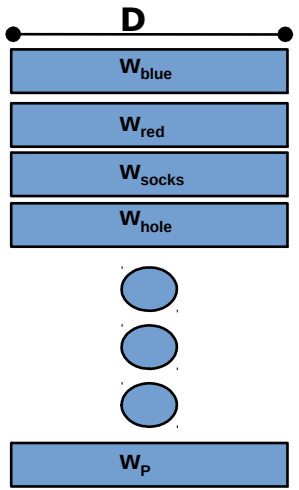


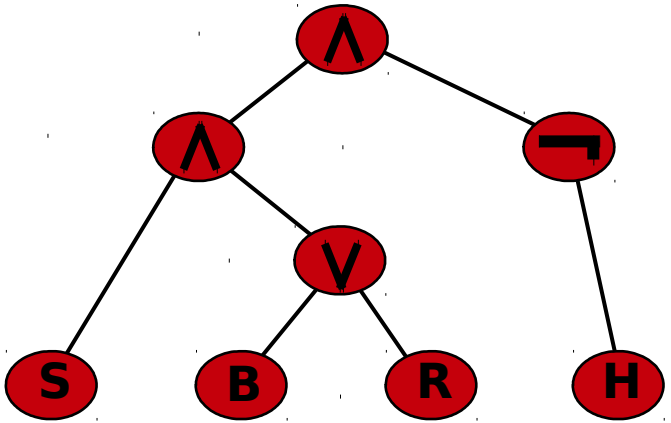
$g^{not}(\bullet): R^D \rightarrow R^D$

$g^{\wedge}(\bullet, \bullet): R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet): R^D \times R^D \rightarrow R^D$

Inference





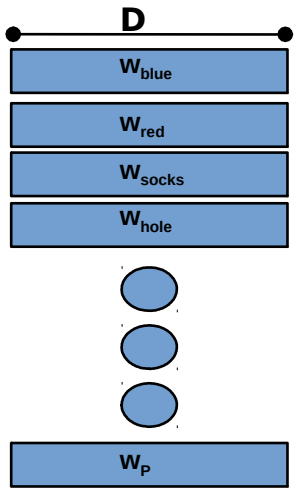
Blue or **R**ed **S**ocks Without **H**oles

$g^{not}(\bullet)$: $R^D \rightarrow R^D$

$g^{\wedge}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$

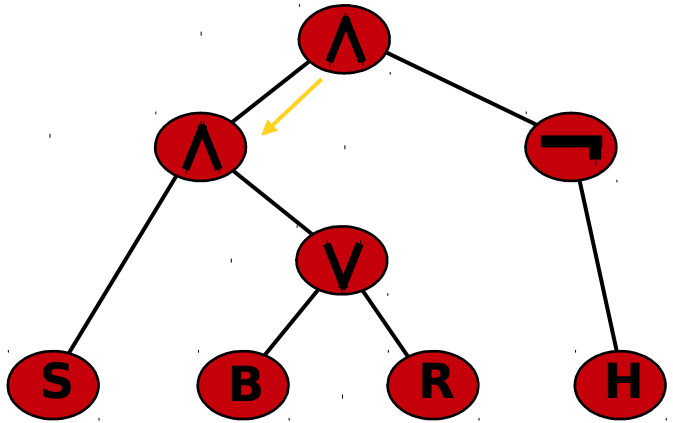
Inference



$g^{not}(\bullet)$: $R^D \rightarrow R^D$

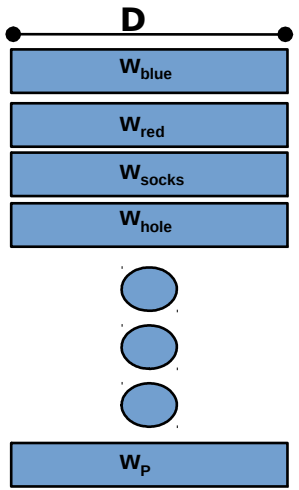
$g^{\wedge}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$



Blue or **R**ed **S**ocks Without **H**oles

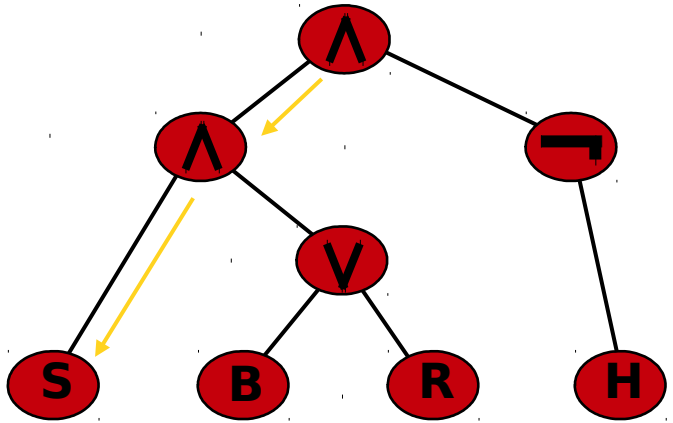
Inference



$g^{not}(\bullet)$: $R^D \rightarrow R^D$

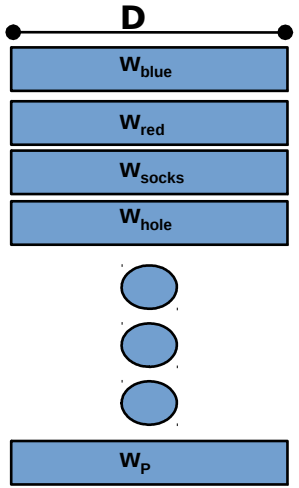
$g^{\wedge}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$



Blue or **R**ed **S**ocks Without **H**oles

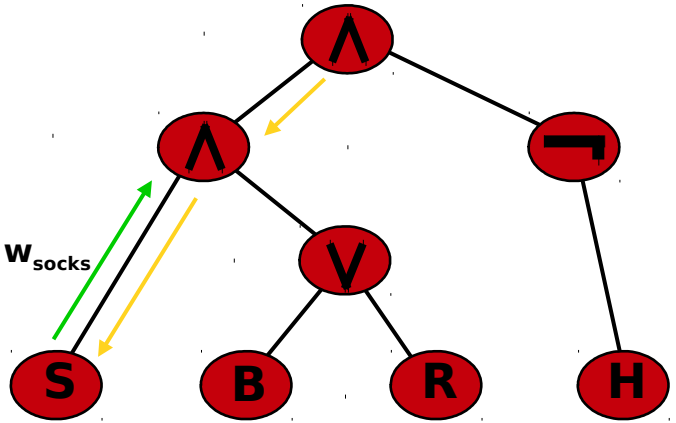
Inference



$g^{not}(\bullet)$: $R^D \rightarrow R^D$

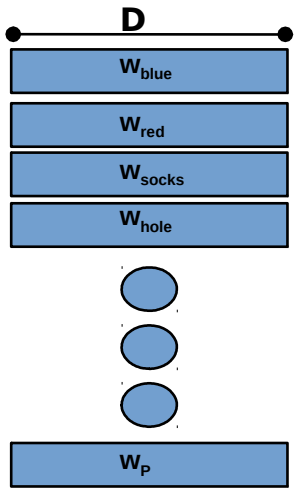
$g^{\wedge}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$



Blue or **R**ed **S**ocks Without **H**oles

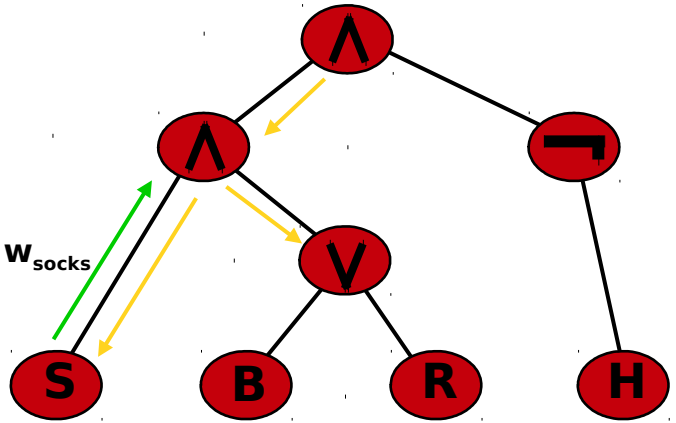
Inference



$g^{not}(\bullet)$: $R^D \rightarrow R^D$

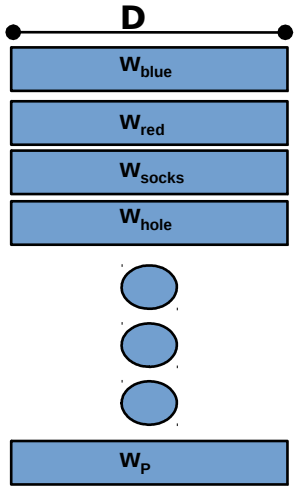
$g^{\wedge}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$



Blue or **R**ed **S**ocks Without **H**oles

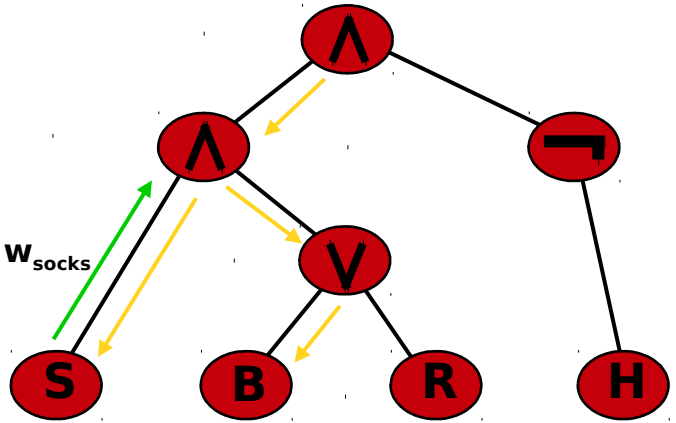
Inference



$g^{not}(\bullet)$: $R^D \rightarrow R^D$

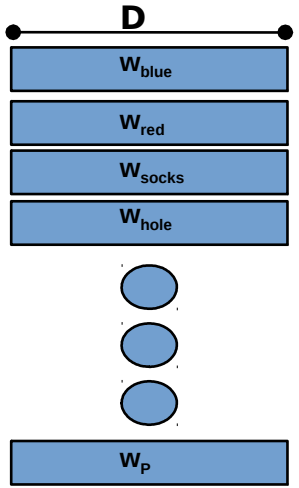
$g^{\wedge}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$



Blue or **R**ed **S**ocks Without **H**oles

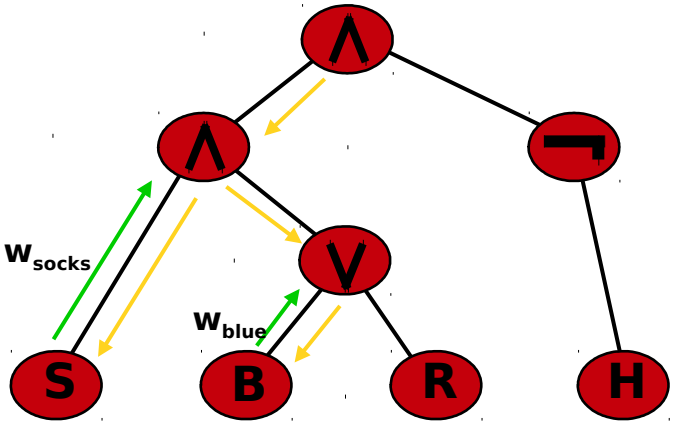
Inference



$g^{not}(\bullet)$: $R^D \rightarrow R^D$

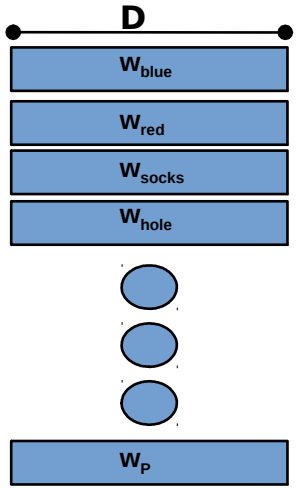
$g^{\wedge}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$



Blue or **R**ed **S**ocks Without **H**oles

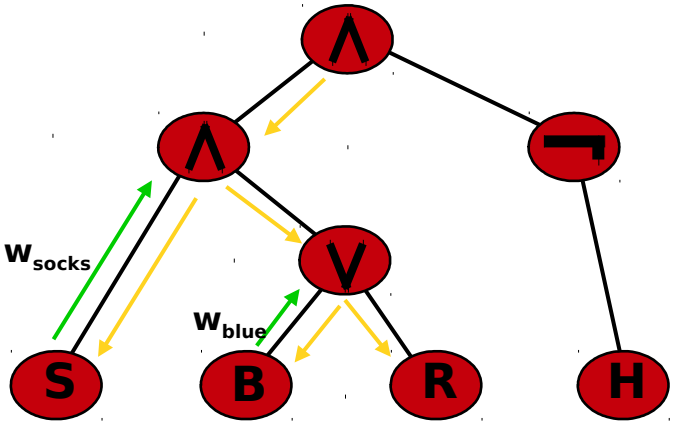
Inference



$g^{not}(\bullet)$: $R^D \rightarrow R^D$

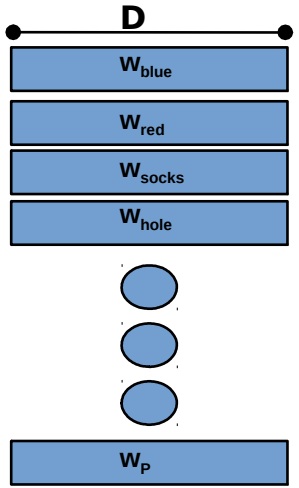
$g^{\wedge}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$



Blue or **R**ed **S**ocks Without **H**oles

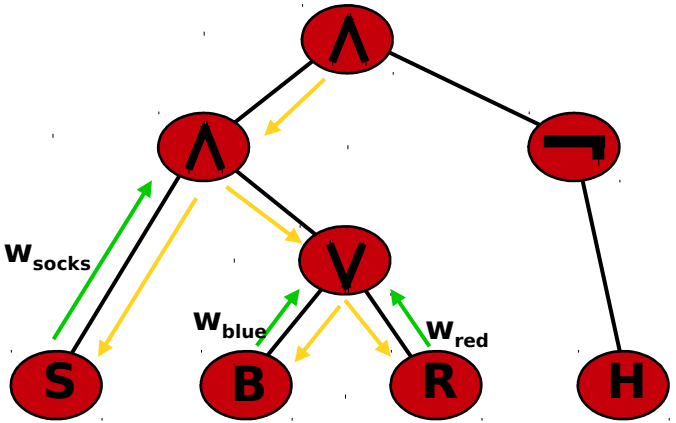
Inference



$g^{not}(\bullet)$: $R^D \rightarrow R^D$

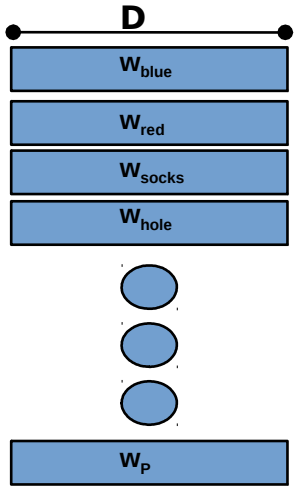
$g^{\wedge}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$



Blue or **R**ed **S**ocks Without **H**oles

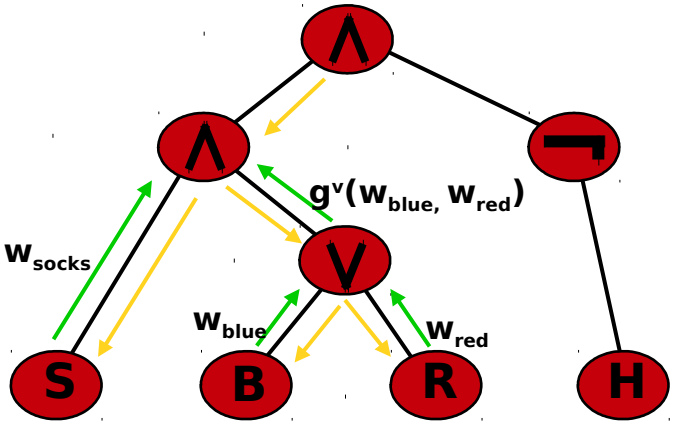
Inference



$g^{not}(\bullet)$: $R^D \rightarrow R^D$

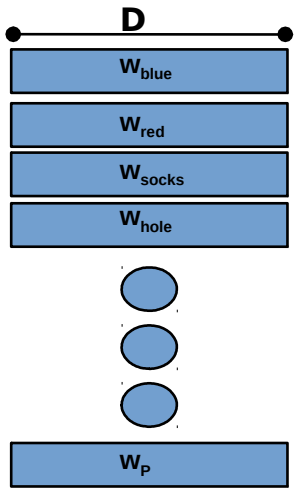
$g^{\wedge}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$



Blue or **R**ed **S**ocks Without **H**oles

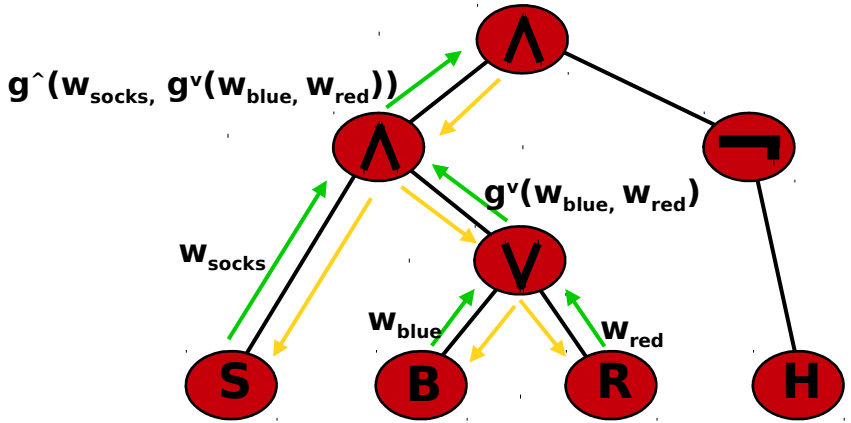
Inference



$g^{not}(\bullet)$: $R^D \rightarrow R^D$

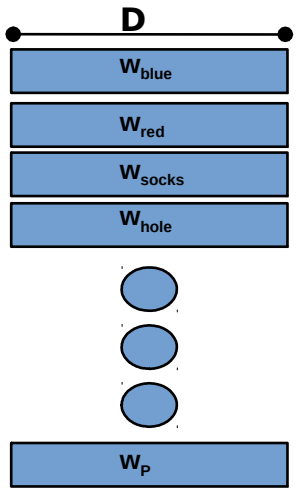
$g^{\wedge}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$



Blue or **R**ed **S**ocks Without **H**oles

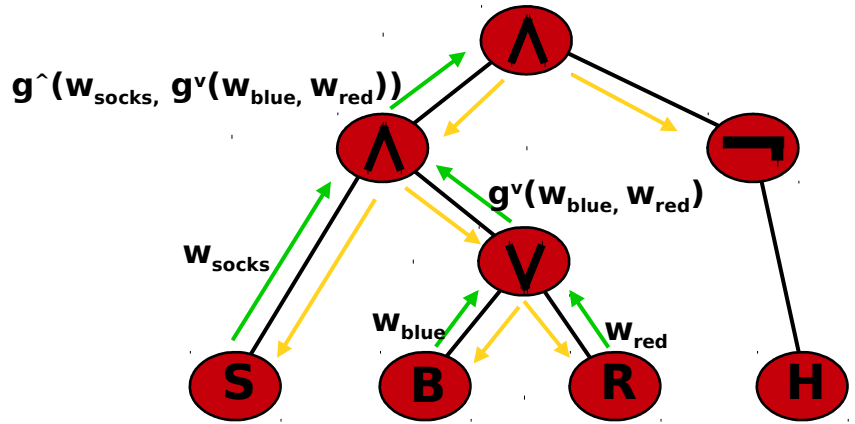
Inference



$g^{not}(\bullet)$: $R^D \rightarrow R^D$

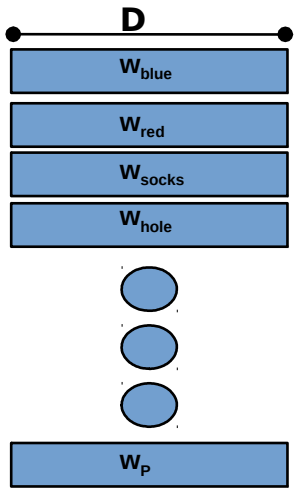
$g^{\wedge}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$



Blue or **R**ed **S**ocks Without **H**oles

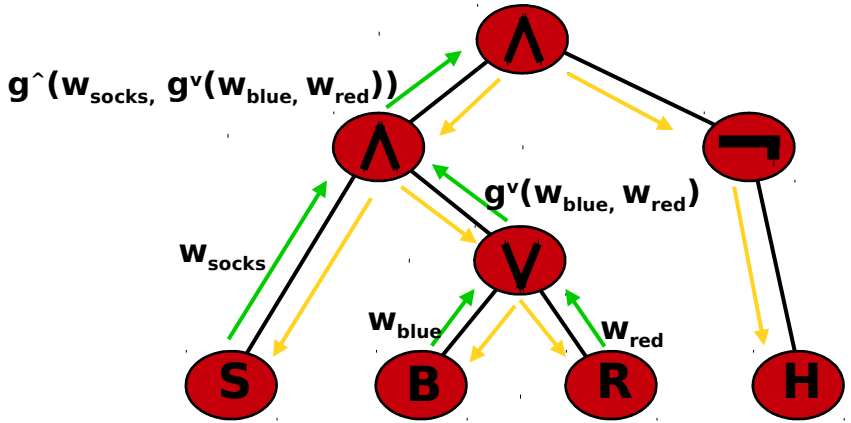
Inference



$g^{not}(\bullet)$: $R^D \rightarrow R^D$

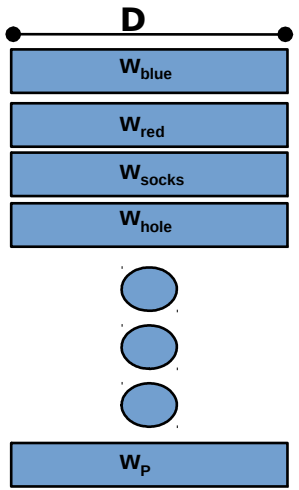
$g^{\wedge}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$



Blue or **R**ed **S**ocks Without **H**oles

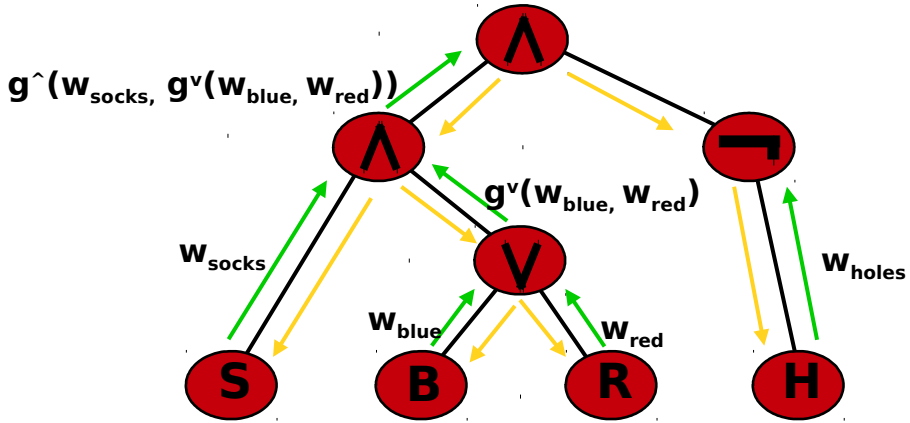
Inference



$g^{not}(\bullet)$: $R^D \rightarrow R^D$

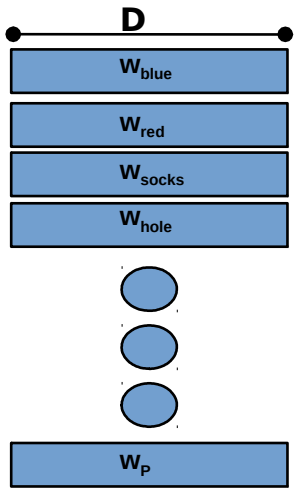
$g^{\wedge}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$



Blue or **R**ed **S**ocks Without **H**oles

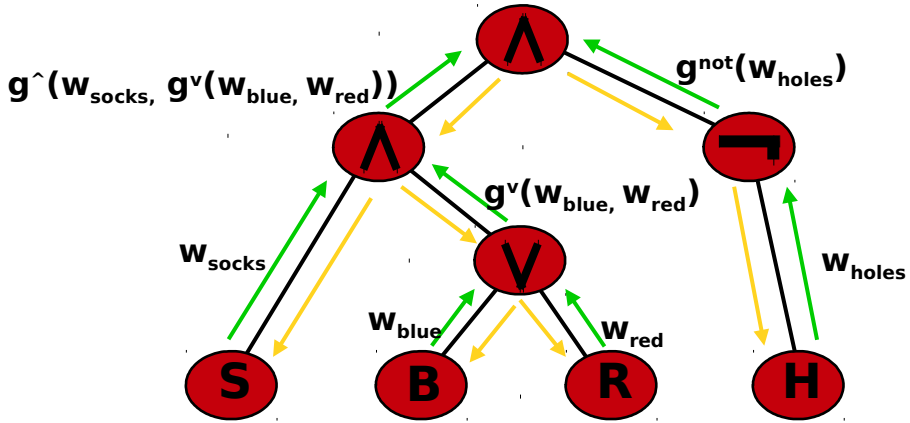
Inference



$g^{not}(\bullet)$: $R^D \rightarrow R^D$

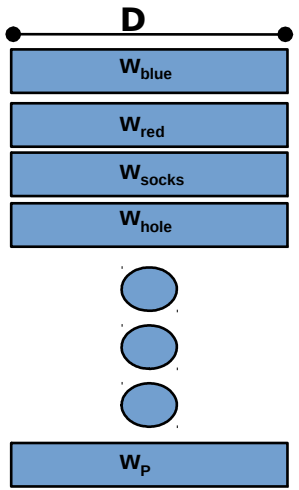
$g^{\wedge}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$



Blue or **R**ed **S**ocks Without **H**oles

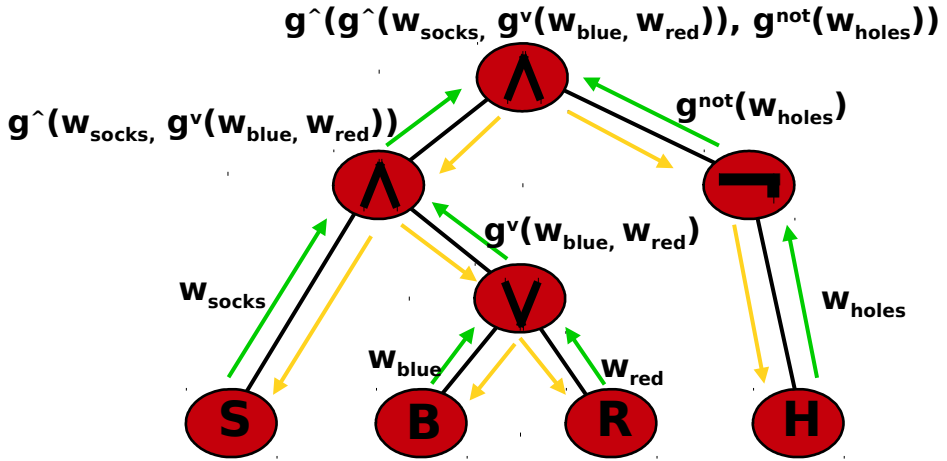
Inference



$g^{not}(\bullet)$: $R^D \rightarrow R^D$

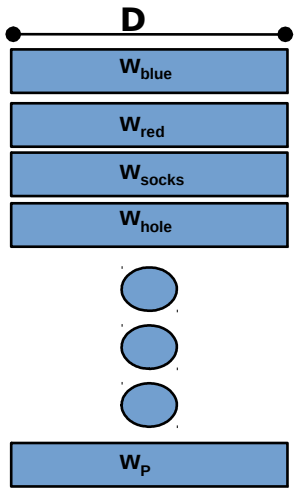
$g^{\wedge}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$

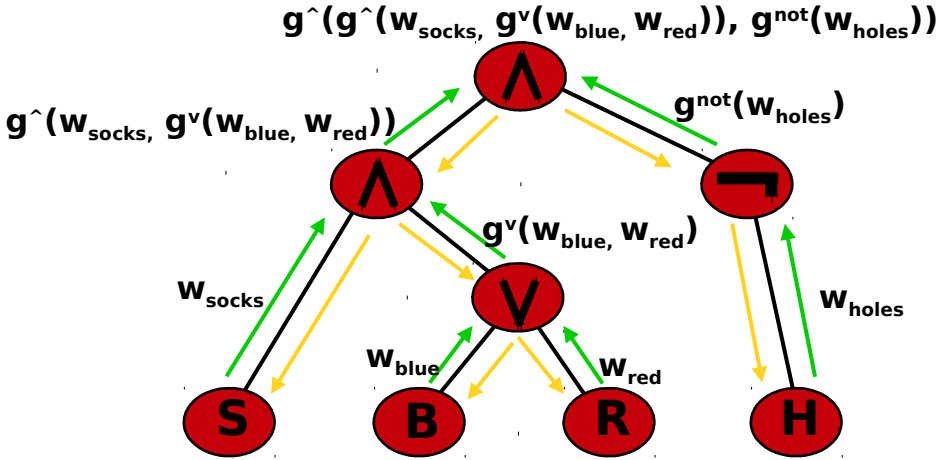
$g^{v}(\bullet, \bullet)$: $R^D \times R^D \rightarrow R^D$



Blue or **R**ed **S**ocks Without **H**oles

Inference





Blue or **R**ed **S**ocks Without **H**oles

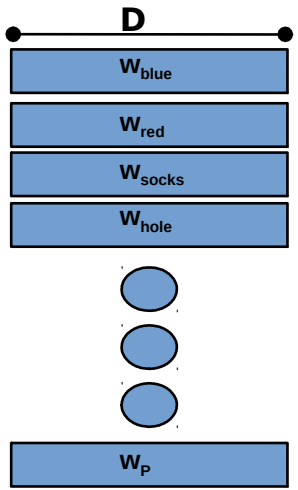
$g^{not}(\bullet): R^D \rightarrow R^D$

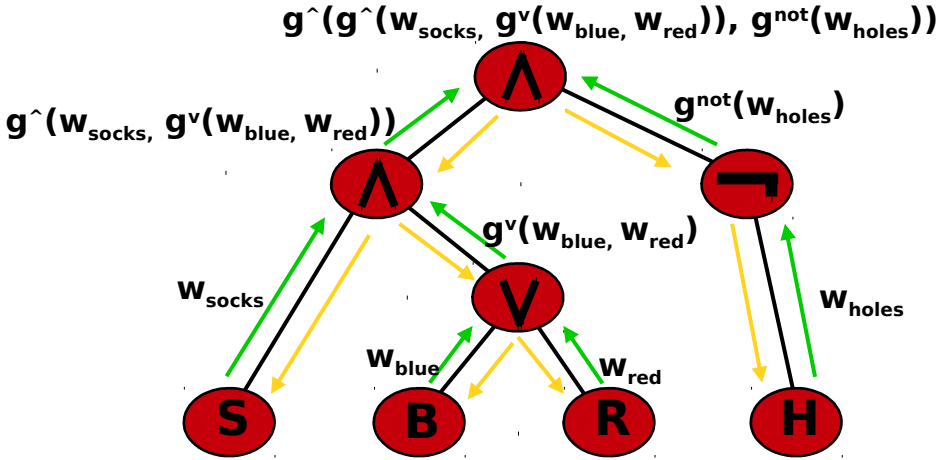
$g^{\hat{}}(\bullet, \bullet): R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet): R^D \times R^D \rightarrow R^D$

$f(e) = g^{\hat{}}(g^{\hat{}}(w_{socks}, g^{v}(w_{blue}, w_{red})), g^{not}(w_{holes}))$

Inference





Blue or **R**ed **S**ocks Without **H**oles

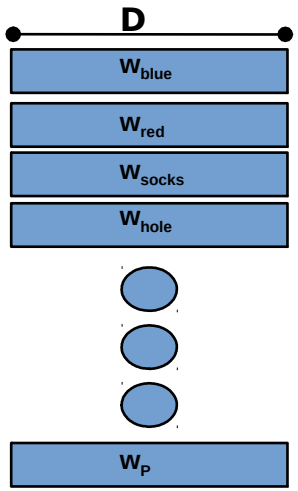
$g^{not}(\bullet): R^D \rightarrow R^D$

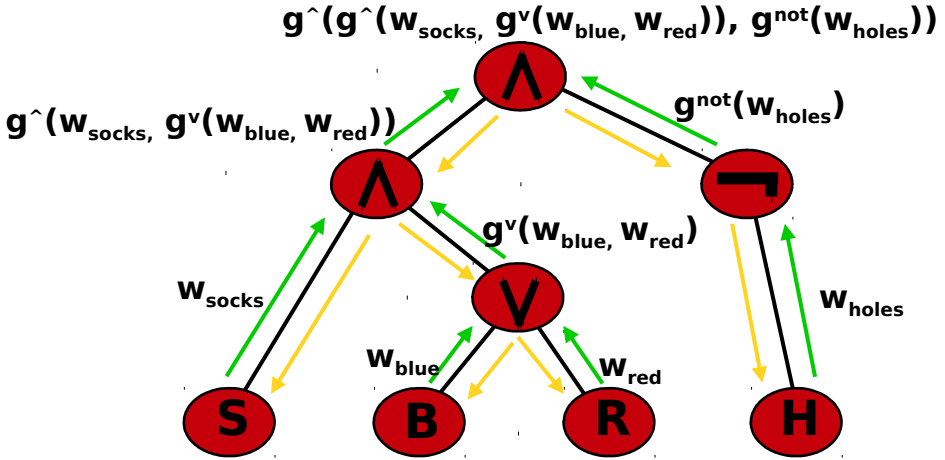
$g^{\hat{}}(\bullet, \bullet): R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet): R^D \times R^D \rightarrow R^D$

$$f(e) = g^{\hat{}}(g^{\hat{}}(w_{socks}, g^{v}(w_{blue}, w_{red})), g^{not}(w_{holes}))$$
$$= w_e \in R^D$$

Inference





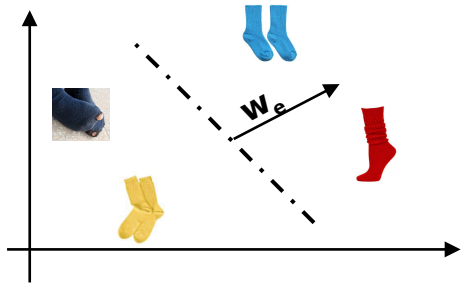
Blue or **R**ed **S**ocks Without **H**oles

$g^{not}(\bullet): R^D \rightarrow R^D$

$g^{\hat{}}(\bullet, \bullet): R^D \times R^D \rightarrow R^D$

$g^{v}(\bullet, \bullet): R^D \times R^D \rightarrow R^D$

$$f(e) = g^{\hat{}}(g^{\hat{}}(w_{socks}, g^{v}(w_{blue}, w_{red})), g^{not}(w_{holes}))$$
$$= w_e \in R^D$$



Simplifying ...

We model our function as a set of composition functions and simplify them using simple analytical relations and De Morgan’s laws.

$$g_\theta^{\wedge}(w_a, w_b) = \text{Neural Network}(w_a, w_b)$$

$$g^{\neg}(w) = -w$$

$$g^{\vee}(w_a, w_b) = g^{\neg}\left(g^{\wedge}\left(g^{\neg}(w_a), g^{\neg}(w_b)\right)\right)$$

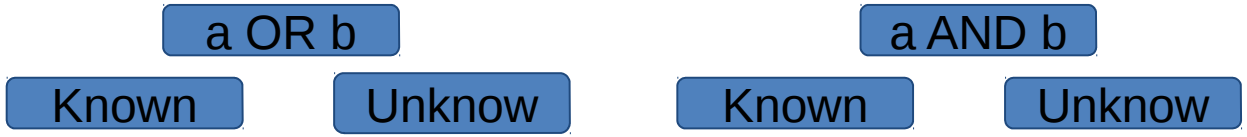
Experiments - Binary Expressions

Experiments - Binary Expressions

a OR b

a AND b

Experiments - Binary Expressions



Experiments - Binary Expressions

a OR b

a AND b

Known

Unknow

Known

Unknow

Table 1. Evaluating known/unk nd conjunctive expressions on the CU t.

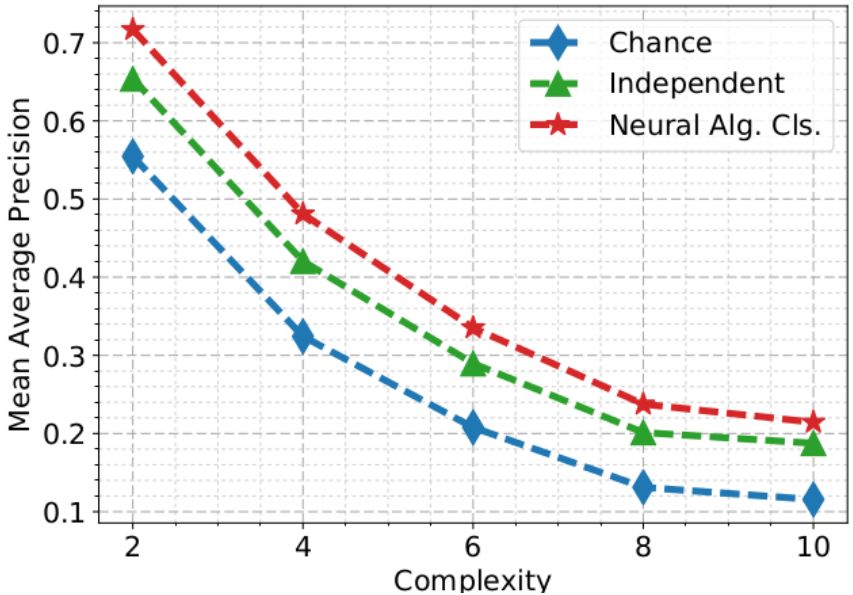
	Disjunctive Expressions						Conjunctive Expressions					
Metrics	MAP	AUC	EER	MAP	AUC	EER	MAP	AUC	EER	MAP	AUC	EER
Chance	39.70	50.00	50.0	40.60	50.00	50.0	4.55	50.0	50.0	4.59	50.0	50.0
Supervised	65.25	74.76	31.58	-	-	-	22.87	78.02	29.69	-	-	-
Independent	58.73	68.39	36.76	60.66	69.28	36.10	17.23	77.22	29.94	19.16	78.00	29.28
Neural Alg. Classifiers	**70.10**	**77.36**	**29.44**	**71.18**	**77.76**	**29.04**	**23.09**	**81.54**	**26.36**	**23.87**	**81.98**	**25.85**

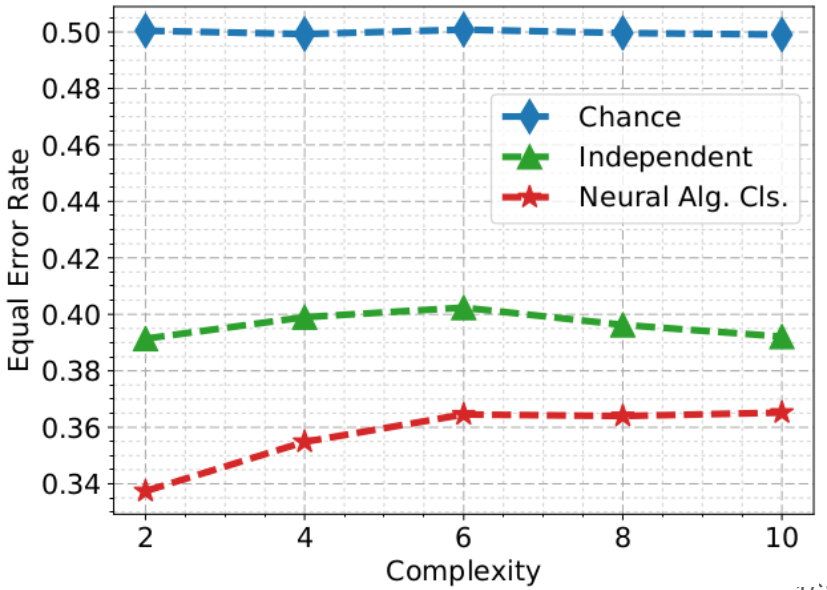
Table 2. Evaluating known/unknown disjunctive and conjunctive expressions on the AwA2 dataset.

	Disjunctive Expressions						Conjunctive Expressions					
	Known Exp.			Unknown Exp.			Known Exp.			Unknown Exp.		
Metrics	MAP	AUC	EER	MAP	AUC	EER	MAP	AUC	EER	MAP	AUC	EER
Chance	53.19	50.0	50.0	53.04	50.0	50.0	18.77	50.0	50.0	21.17	50.0	50.0
Supervised	97.47	97.20	8.13	-	-	-	94.90	98.53	6.00	-	-	-
Independent	97.28	97.12	8.70	97.86	97.58	6.77	93.95	98.13	6.80	93.90	97.87	7.36
Neural Alg. Classifiers	**98.84**	**98.67**	**5.84**	**99.05**	**98.91**	**5.24**	**95.95**	**98.79**	**5.29**	**96.50**	**98.81**	**5.34**

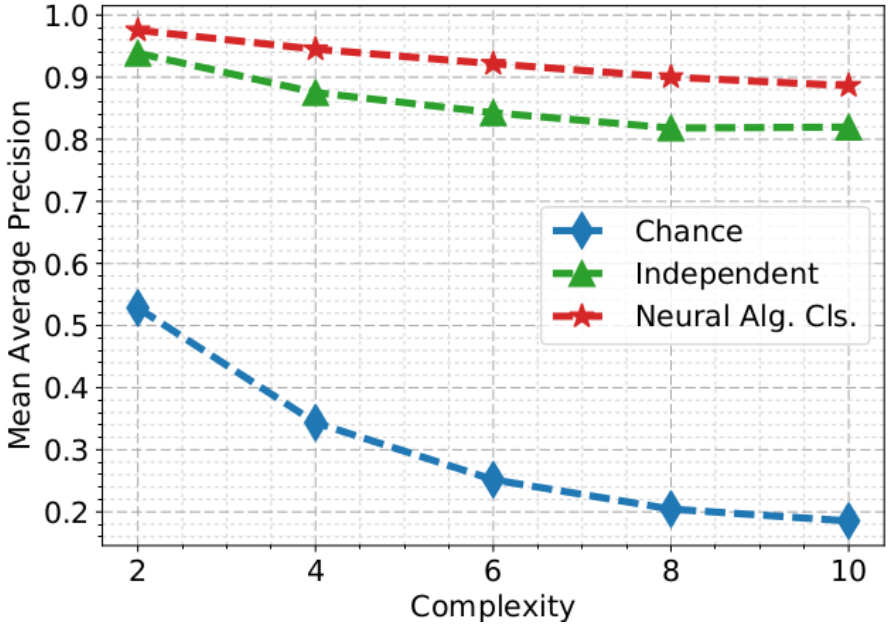
Experiments - Complex Expressions - CUB200

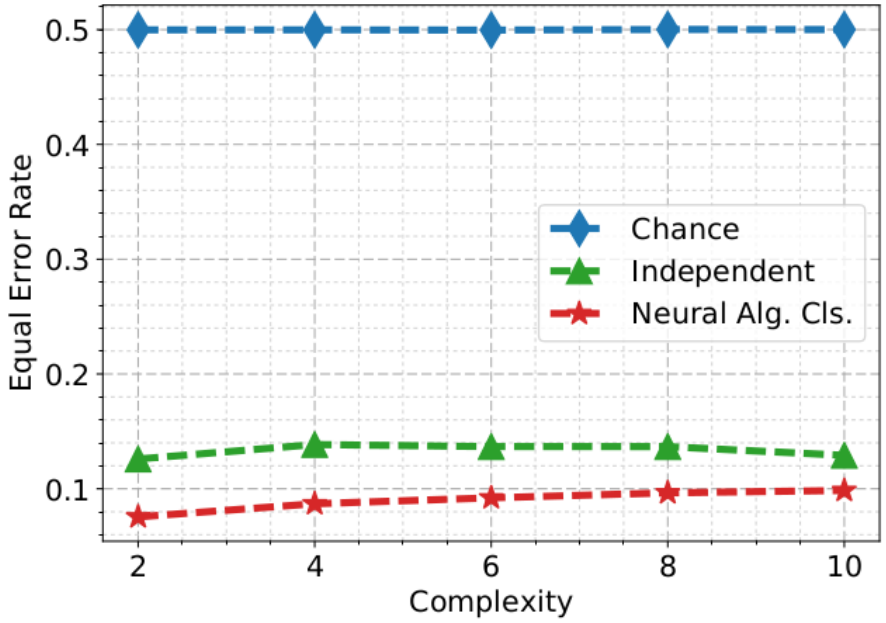
Complex Unknown Expressions in Conjunctive Normal Form (CNF):
$(p_1 \vee q_1) \wedge (p_2 \vee q_2) \wedge \ldots. (p_c \vee q_c)$





Experiments - Complex Expressions - AWA2





Qualitative Experiments

Qualitative Experiments

Birds with crown and breast of **the same** color (e.g., blue, yellow, or red.)

Qualitative Experiments

Birds with crown and breast of **the same** color (e.g., blue, yellow, or red.)

Birds with crown and breast of **different** color (e.g., blue, yellow, or red.)

Qualitative Experiments

Birds with crown and breast of **the same** color (e.g., blue, yellow, or red.)

Birds with crown and breast of **different** color (e.g., blue, yellow, or red.)

Qualitative Experiments

Birds with crown and breast of **the same** color (e.g., blue, yellow, or red.)

Birds with crown and breast of **different** color (e.g., blue, yellow, or red.)

Qualitative Experiments

Birds with crown and breast of **the same** color (e.g., blue, yellow, or red.)

Birds with crown and breast of **different** color (e.g., blue, yellow, or red.)

Big and fast animals that are not hunters: *(B AND F) AND (NOT H) = (NOT (S OR SL)) AND (NOT H)*

Qualitative Experiments

Birds with crown and breast of **the same** color (e.g., blue, yellow, or red.)

Birds with crown and breast of **different** color (e.g., blue, yellow, or red.)

Big and fast animals that are not hunters: *(B AND F) AND (NOT H) = (NOT (S OR SL)) AND (NOT H)*

Qualitative Experiments

Birds with crown and breast of **the same** color (e.g., blue, yellow, or red.)

Birds with crown and breast of **different** color (e.g., blue, yellow, or red.)

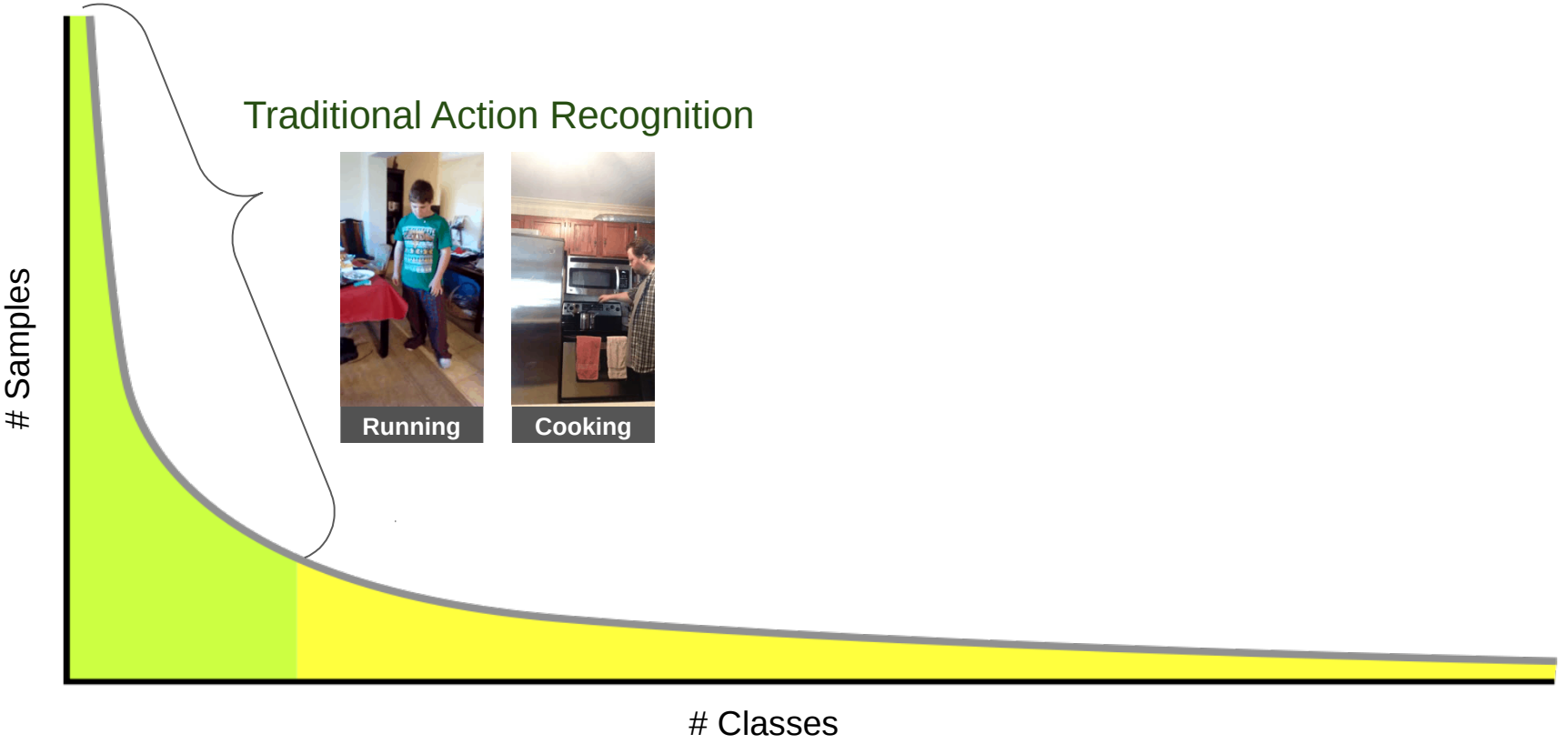
Big and fast animals that are not hunters: *(B AND F) AND (NOT H) = (NOT (S OR SL)) AND (NOT H)*

Extending Existing Models

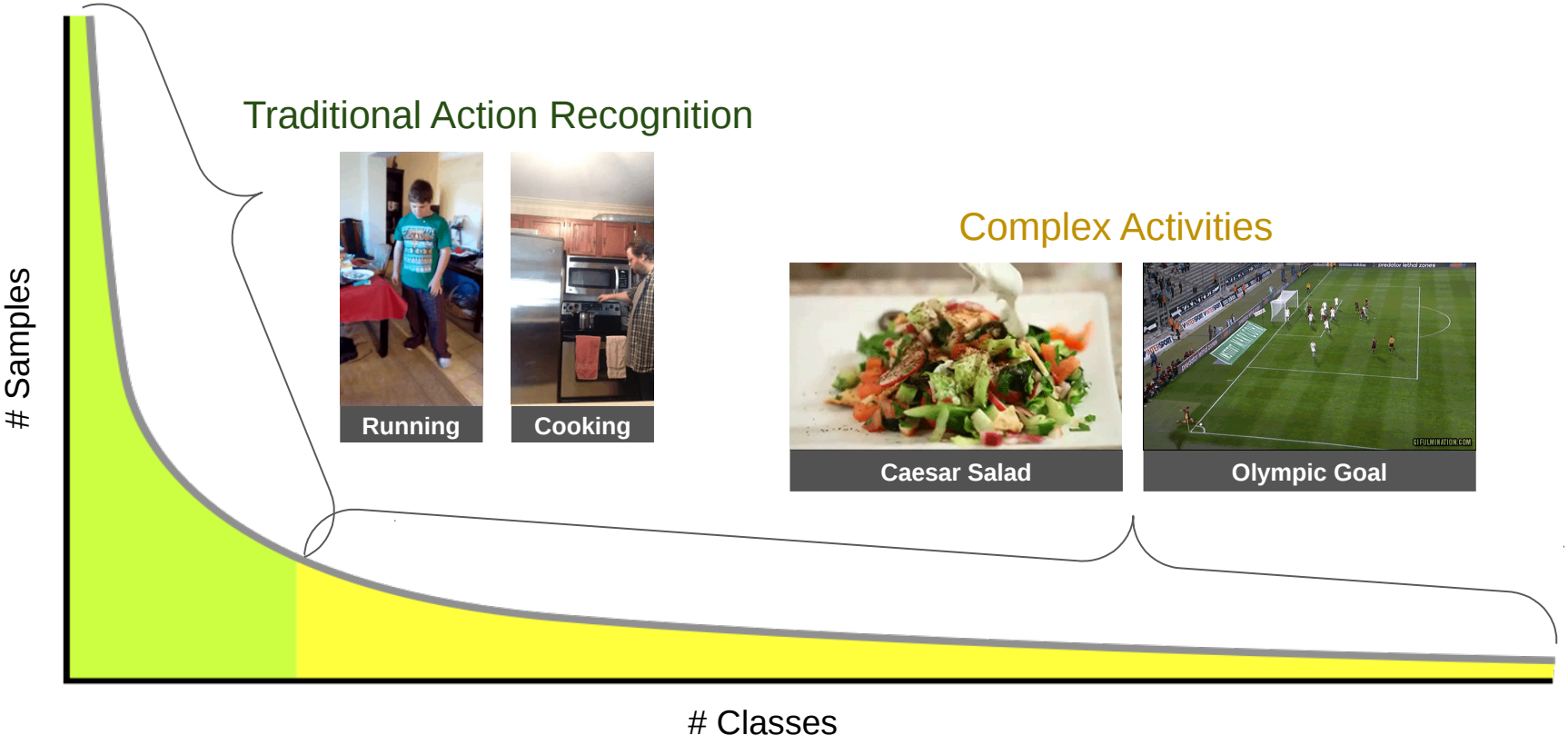
[Inferring Rich Compositional Activities in Videos. **Rodrigo Santa Cruz**, Dylan Campbell, Basura Fernando, Anoop Cherian, Stephen Gould. In IEEE international conference on computer vision (ICCV), 2019. **(Under Review)**]

The long tail of complex activities

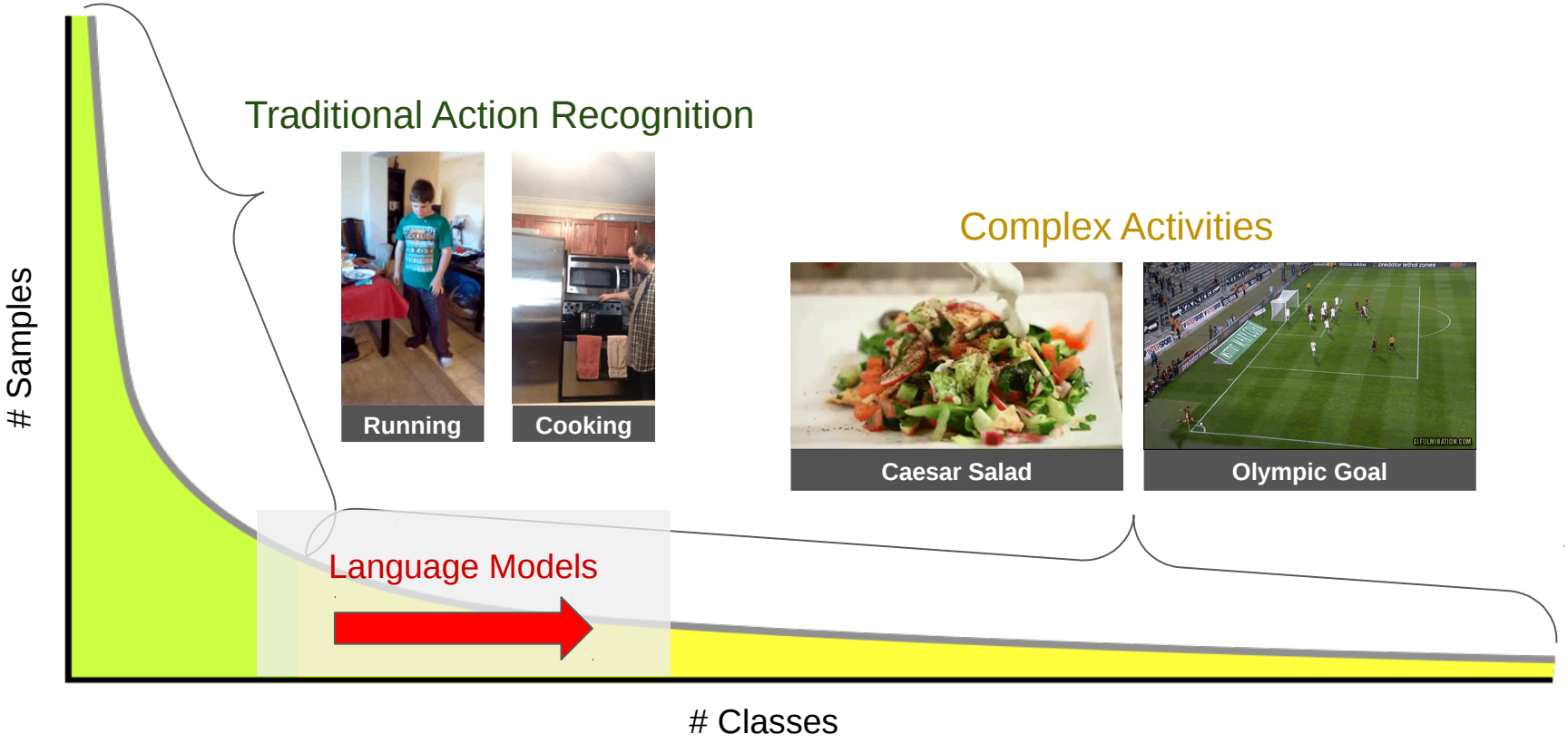
The long tail of complex activities



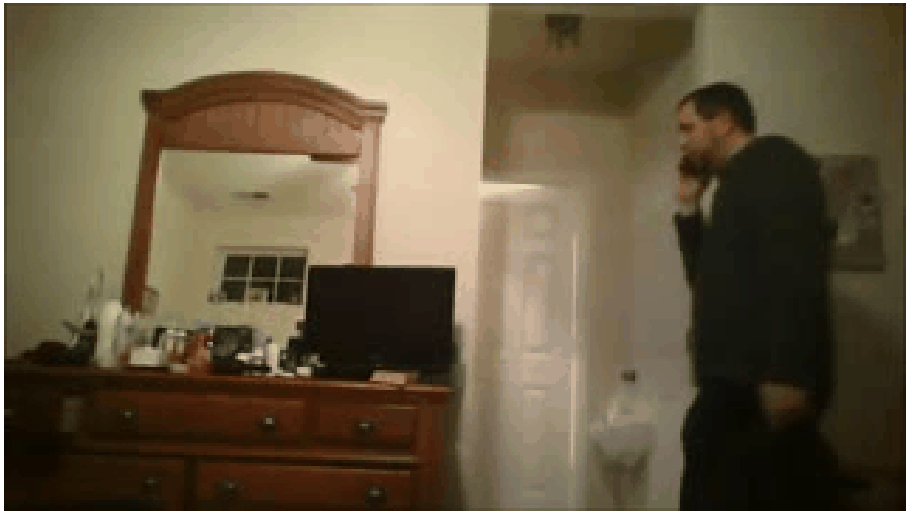
The long tail of complex activities



The long tail of complex activities



Activity Recognition from Natural Language



Someone . . .

“. . . is talking on the phone, dressing a jacket and brushing hair.”

“ . . . is talking on the phone and holding a jacket, then he dresses it and brushes his hair.”

“. . . is talking on the phone while dressing a jacket and brushing hair.”

Rich Compositional Activity Recognition

We propose inference framework which can recognize complex activities expressed as regular expressions of simpler actions.

Rich Compositional Activity Recognition

We propose inference framework which can recognize complex activities expressed as regular expressions of simpler actions.

Simple actions:
TP = talks on the phone; HJ = Holding a jacket;
D = Dressing; BH = Brushing Hair;

Rich Compositional Activity Recognition

We propose inference framework which can recognize complex activities expressed as regular expressions of simpler actions.

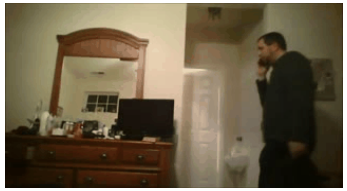
Simple actions:
TP = talks on the phone; HJ = Holding a jacket;
D = Dressing; BH = Brushing Hair;

(TP, HJ)(TP, D)(TP, BH)

Rich Compositional Activity Recognition

We propose inference framework which can recognize complex activities expressed as regular expressions of simpler actions.

Simple actions:
TP = talks on the phone; HJ = Holding a jacket;
D = Dressing; BH = Brushing Hair;



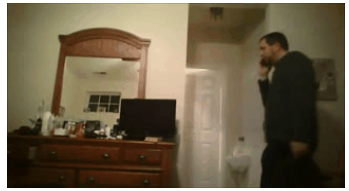
(TP, HJ)(TP, D)(TP, BH)

Rich Compositional Activity Recognition

We propose inference framework which can recognize complex activities expressed as regular expressions of simpler actions.

Simple actions:
TP = talks on the phone; HJ = Holding a jacket;
D = Dressing; BH = Brushing Hair;

(TP, HJ)(TP, D)(TP, BH)

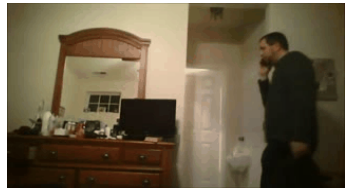


Rich Compositional Activity Recognition

We propose inference framework which can recognize complex activities expressed as regular expressions of simpler actions.

Simple actions:
TP = talks on the phone; HJ = Holding a jacket;
D = Dressing; BH = Brushing Hair;

(TP, HJ)(TP, D)(TP, BH)



It allows to recognize new, specific instances, and groups of activities without additional annotation effort in a unambiguously fashion.

Problem Formulation

Problem Formulation

Describe complex activities by regular expressions of subset of primitive actions (symbols):

Problem Formulation

Describe complex activities by regular expressions of subset of primitive actions (symbols):

Primitives
$\mathcal{A} = \{a_i\}_{i=1}^{M}$

Symbols
$w \in \mathcal{P}(\mathcal{A})$

Regex Operators
$\mathcal{O} = \{\succ, |, \star\}$

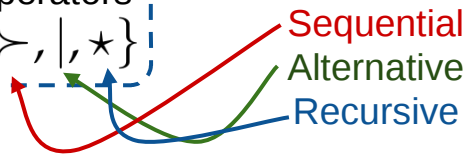
Problem Formulation

Describe complex activities by regular expressions of subset of primitive actions (symbols):

Primitives
$\mathcal{A} = \{a_i\}_{i=1}^{M}$

Symbols
$w \in \mathcal{P}(\mathcal{A})$

Regex Operators
$\mathcal{O} = \{\succ, |, \star\}$



Problem Formulation

Describe complex activities by regular expressions of subset of primitive actions (symbols):

Primitives
$\mathcal{A} = \{a_i\}_{i=1}^{M}$

Symbols
$w \in \mathcal{P}(\mathcal{A})$

Regex Operators
$\mathcal{O} = \{\succ, |, \star\}$

Sequential
Alternative
Recursive

Ex: “driving (a_d) and talking on the phone (a_{tc}) or to someone (a_{ts}) repeatedly just after he got in the car (a_{gc})”

Problem Formulation

Describe complex activities by regular expressions of subset of primitive actions (symbols):

Primitives
$\mathcal{A} = \{a_i\}_{i=1}^{M}$

Symbols
$w \in \mathcal{P}(\mathcal{A})$

Regex Operators
$\mathcal{O} = \{\succ, |, \star\}$

Sequential
Alternative
Recursive

Ex: “driving (a_d) and talking on the phone (a_{tc}) or to someone (a_{ts}) repeatedly just after he got in the car (a_{gc})”

$$a_{gc} \succ (\{a_d,\ a_{tc}\} \mid \{a_d,\ a_{ts}\})^{\star}$$

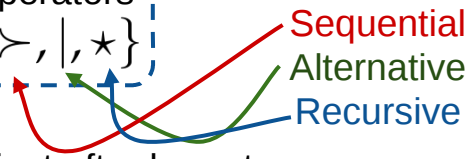
Problem Formulation

Describe complex activities by regular expressions of subset of primitive actions (symbols):

Primitives
$\mathcal{A} = \{a_i\}_{i=1}^{M}$

Symbols
$w \in \mathcal{P}(\mathcal{A})$

Regex Operators
$\mathcal{O} = \{\succ, |, \star\}$



Ex: “driving (a_d) and talking on the phone (a_{tc}) or to someone (a_{ts}) repeatedly just after he got in the car (a_{gc})”

$$a_{gc} \succ (\{a_d,\ a_{tc}\} \mid \{a_d,\ a_{ts}\})^{\star}$$

Then, our goal is to model a function f that assigns high values to a video v if it depicts the action pattern described by the regular expression r and low values otherwise.

$$f_r : \mathcal{V} \rightarrow [0,1]$$

Deterministic Baseline

1. Compile a deterministic finite automaton (DFA) to recognise a given action pattern;
2. Parse video to a subset of action primitives w(x) by thresholding primitive action classifiers at every frame x;

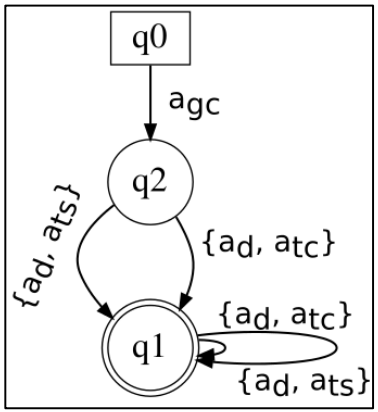
$$w(x) = \{a \in \mathcal{A} \mid p(a|x) \geq \tau\}$$

3. Simulate the DFA with the parsed video;

$$V = [\{a_{gc}\}, \{a_d, a_{tc}\}, \{a_d, a_{tc}\}, \{a_d, a_{tc}\} \ldots]$$

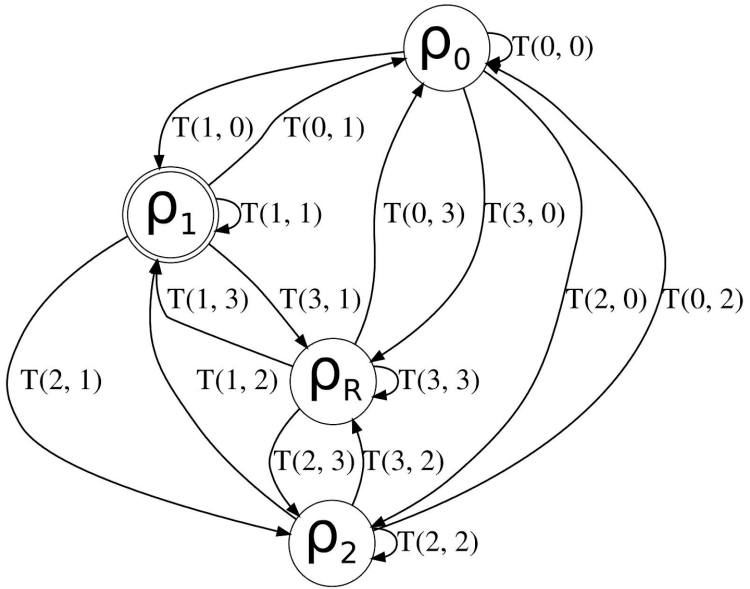
4. Compute the score function.

$$f_r(v) = \frac{dist(q_0, \hat{q})}{dist(q_0, \hat{q}) + \min_{q_f \in \mathcal{F}} dist\left(\hat{q}, q_f\right)}$$



Probabilistic Inference - 1/2

1. Compile the regular expression to a probabilistic automaton (PA) as follow,



$$T_{i,j}(w) = \frac{[\![\delta(i,w) = j]\!] + \alpha}{\sum_{k \in \mathcal{Q}} [\![\delta(i,w) = k]\!] + \alpha|\mathcal{Q}|},$$

$$\boldsymbol{\rho}_i = \frac{[\![q_0 = i]\!] + \alpha}{\sum_{k \in \mathcal{Q}} [\![q_0 = k]\!] + \alpha|\mathcal{Q}|},$$

Probabilistic Inference - 2/2

2. Define an distribution over the power set of action primitives Σ;

$$p(w|x) = \left(\prod_{a \in \mathcal{A}} p(a|x)^{[\![a \in w]\!]} \left(1 - p(a|x)\right)^{\left(1 - [\![a \in w]\!]\right)} \right)^{\gamma}$$

3. We compute the matching probability;

$$P_{U_r}(v) = \left(\boldsymbol{\rho}^{\mathsf{T}} \prod_{i=1}^{|v|} \sum_{w \in \Sigma} T(w) p(w \mid x_i) \right)^{\frac{1}{|v|}} \boldsymbol{f}$$

Experiments - Moving MNIST

We generate videos with different digits appearing patterns expressed by regular expressions of the format,

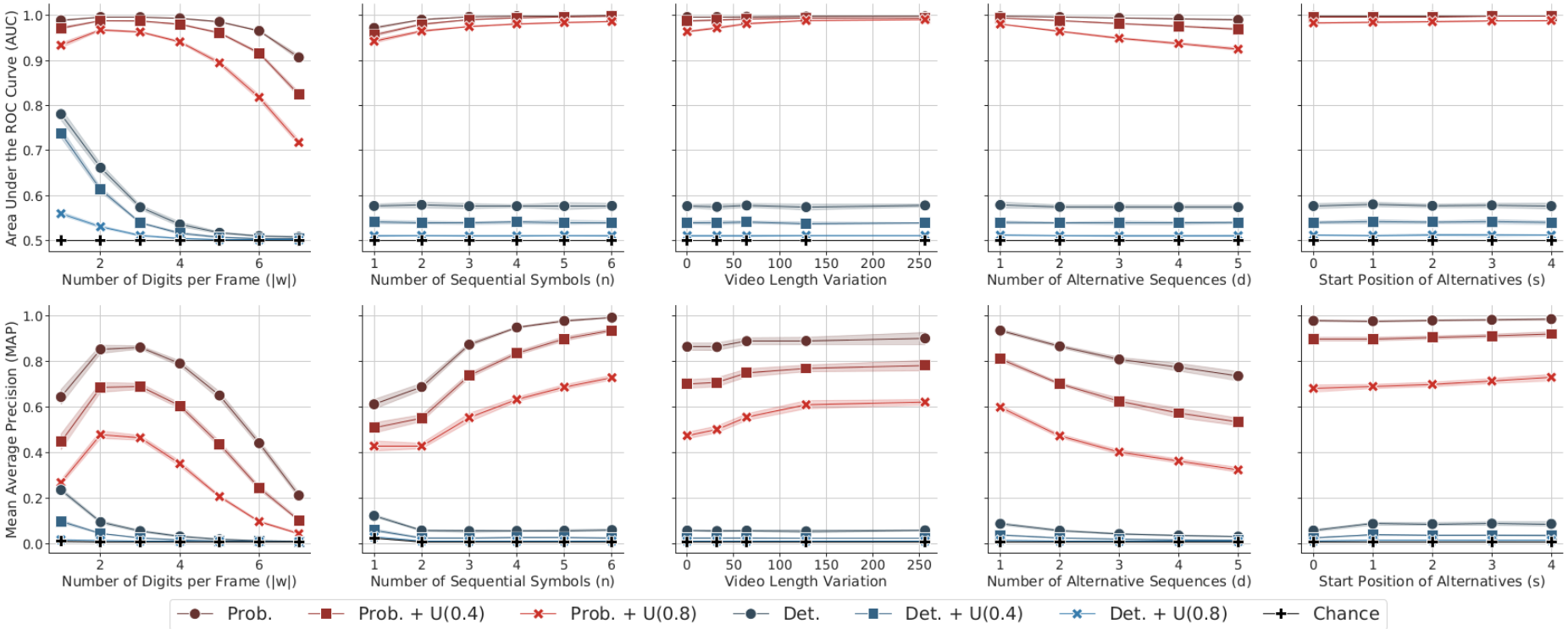
$$w_1^+ \succ \cdots \succ \left(\left(w_s^{1+} \succ \cdots \succ w_n^{1+} \right) \middle| \cdots \middle| \left(w_s^{d+} \succ \cdots \succ w_n^{d+} \right) \right)$$

where the symbols *w* are subsets of the primitives which are the 10 digit classes.

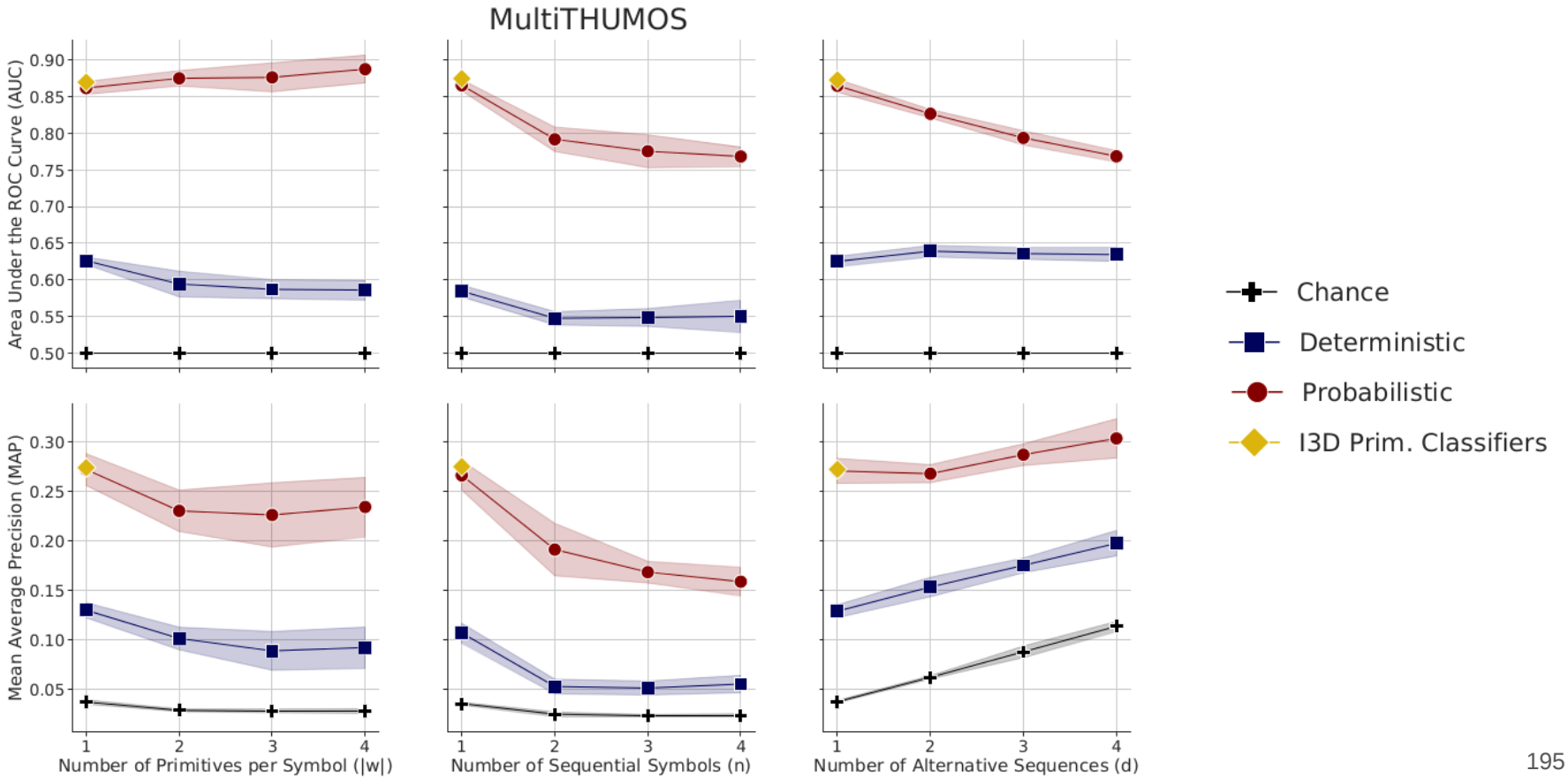
Data generation parameters:

- n: number of sequential patterns
- d: number of alternatives
- s: alternatives start position
- |w|: number of digits appearing simultaneously
- Total number of generated frames

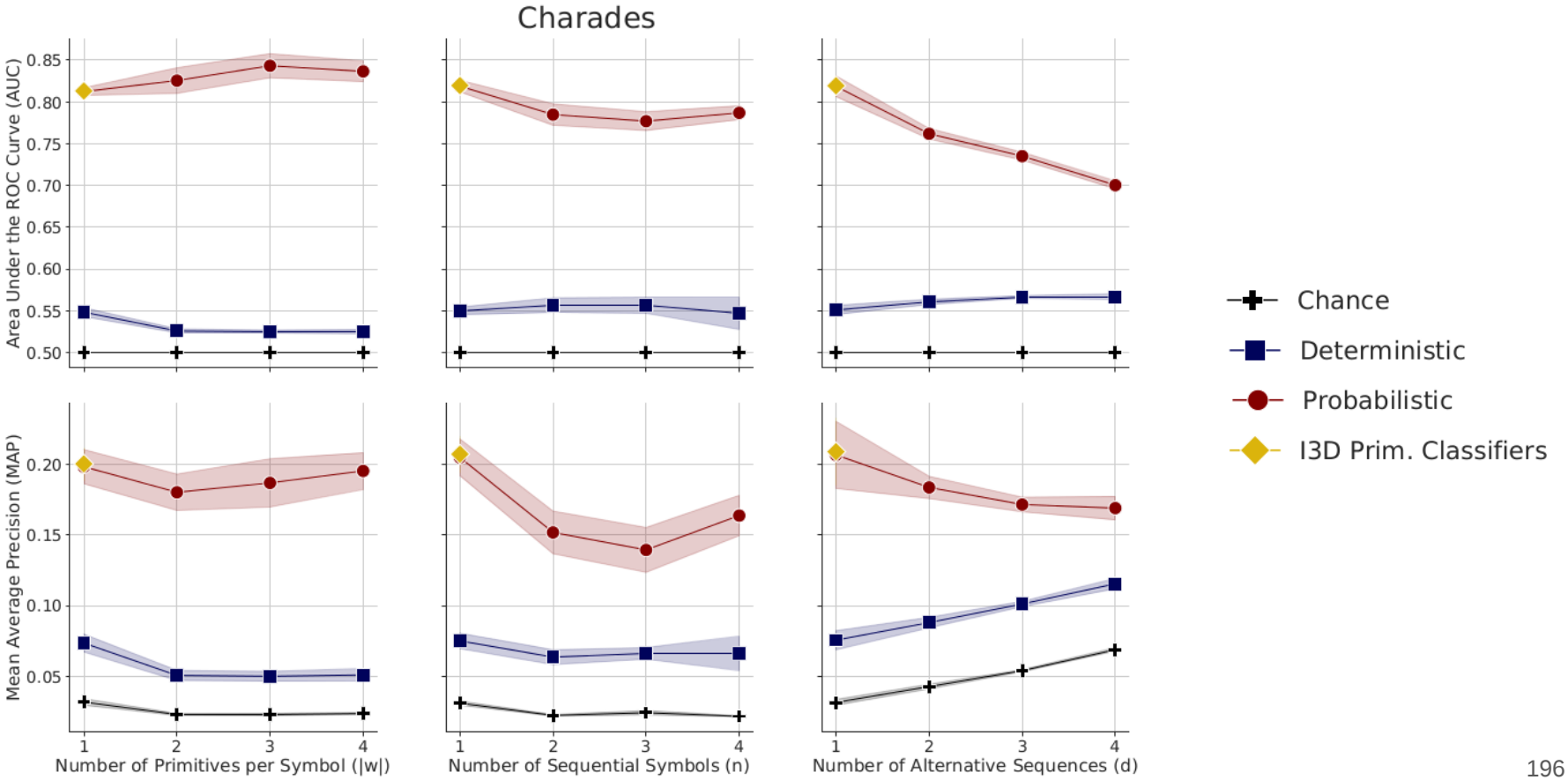
Experiments - Moving MNIST



Experiments - Activity Recognition - MultiTHUMOS



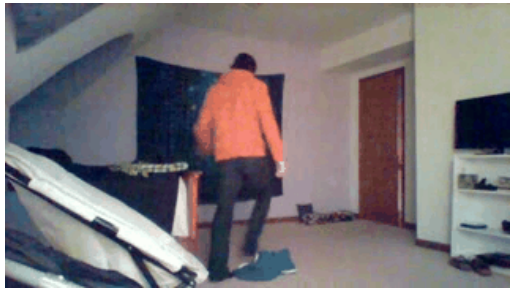
Experiments - Activity Recognition - Charades



Experiments - Qualitative Results

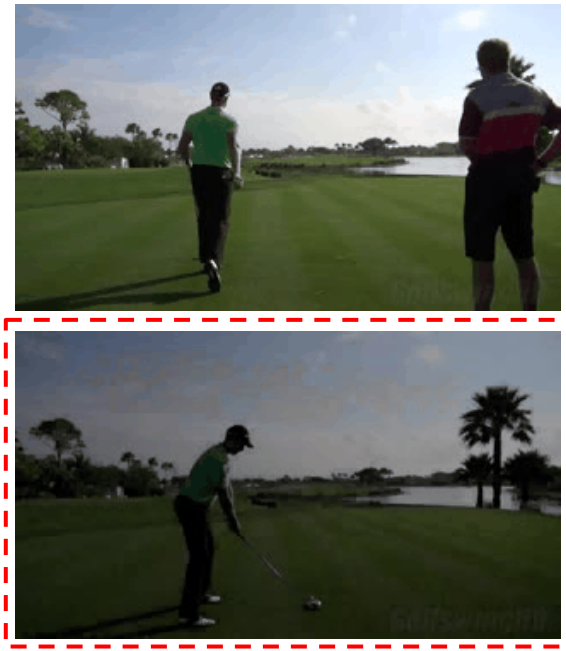
$\{\text{Jump}\}^+ \succ \{\text{Body-Roll}\}^+ \succ \{\text{Body-Bend}\}^+$

$\{\text{WaThDo}\}^+ \succ \{\text{WaThDo, OpDo}\}^+ \succ \{\text{WaThDo, ClDo}\}^+$



Experiments - Qualitative Results - Failures

$\{\text{Stand}\}^+ \succ \{\text{Stand, Throw}\}^+ \succ \{\text{Stand, Golf-Swing}\}^+$



$\{\text{WaThDo, OpDo, GaOnDoKn, HoBa}\}^+ \succ \{\text{WaThDo, ClDo, GaOnDoKn, HoBa}\}^+$

Conclusion

- This thesis...
  - … focuses on reducing the exhaustive human supervision required by the current state-of-the-art models for visual recognition.
  - … presents approaches to overcome the closed world assumption of existing models.
  - … accomplishes its goals by exploring the structure and regularities in the visual world.
  - Applications:
    - Image Ranking.
    - Self-Supervised Representation Learning.
    - Compositional Model for Object Classification.
    - Activity Recognition from Regular Expressions of Primitives.

- Future Work:
  - Visual Permutation Learning Beyond Static Images.
  - Compositional Models Beyond Classification.
  - Modelling Action Correlation, Co-occurrences and Contextuality.

Australian National University

Visual Recognition From Structured Supervision

Rodrigo Santa Cruz and Stephen Gould