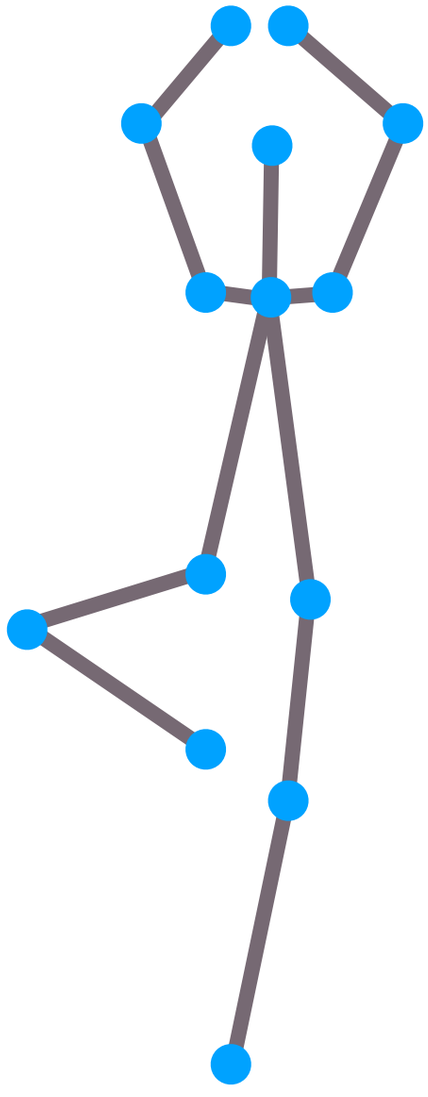
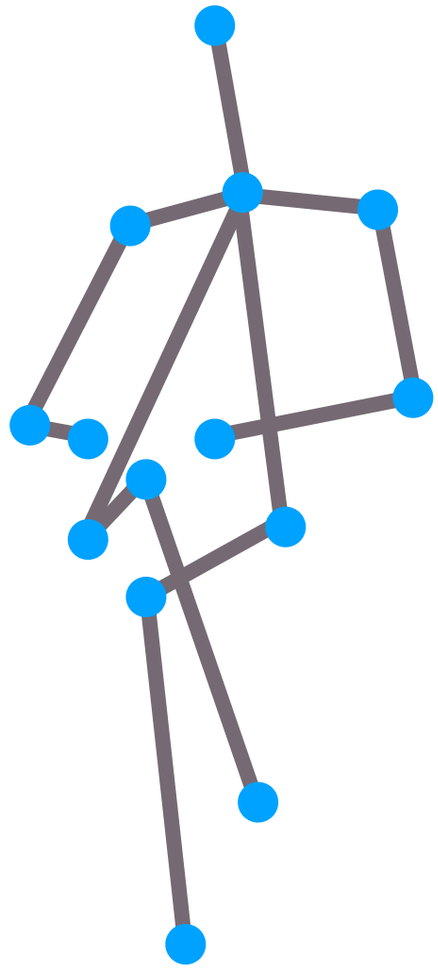
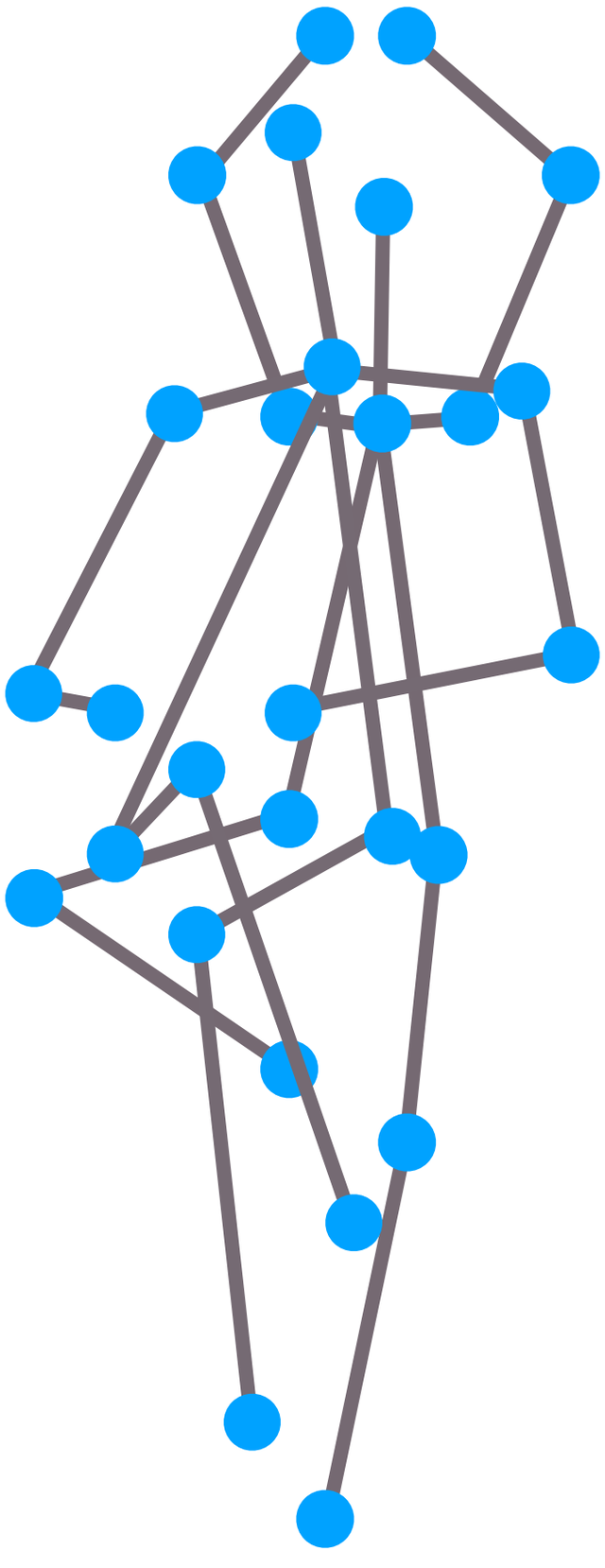
Retrieval

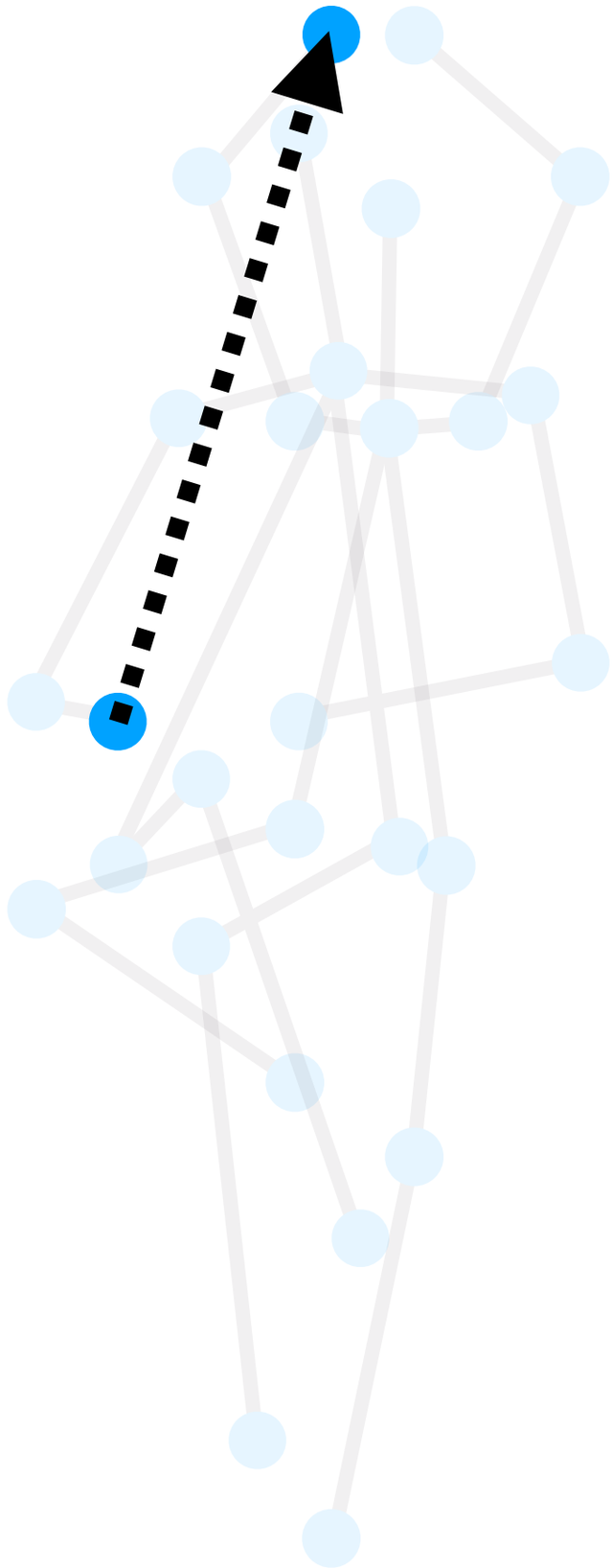
Plan

1. Introduction
2. Background
3. Image Querying
4. Language Querying
5. Language Subset Querying
6. Conclusion









0.1	-
0.13	0.9

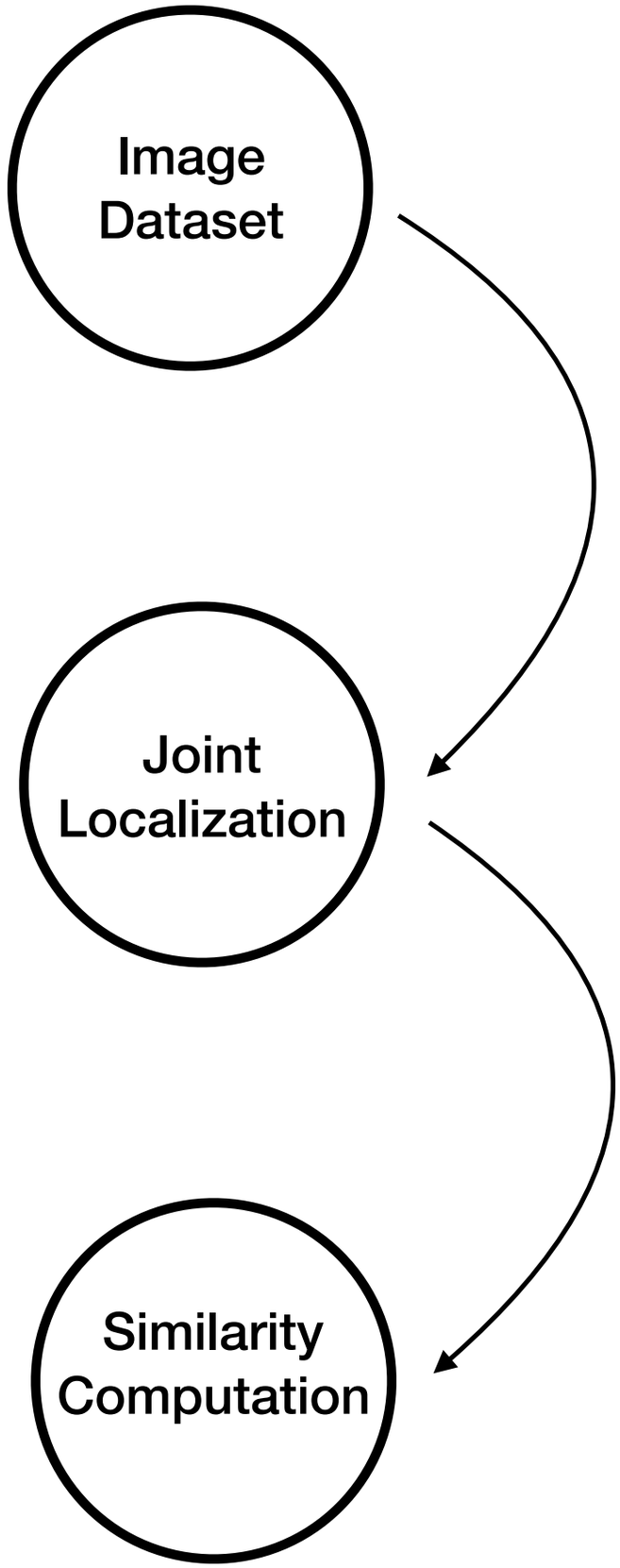


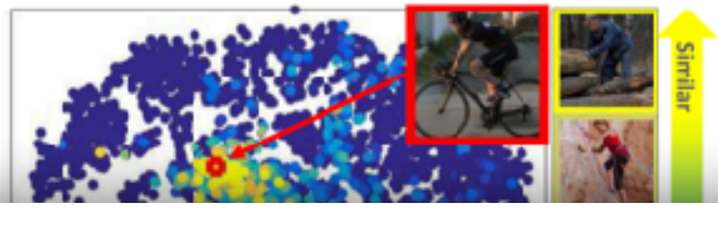
Image
Dataset
Joint
Localization
Similarity
Computation

Thin-Slicing for Pose: Learning to Understand Pose without Explicit Pose Estimation

Suha Kwak Minsu Cho Ivan Laptev
WILLOW Project Team – Inria / École Normale Supérieure, Paris, France

Abstract

We address the problem of learning a pose-aware, compact embedding that projects images with similar human poses to be placed close-by in the embedding space. The embeddi... work, a... real ima... robust ... poses b... backgro... variety ... ding pro... plicit po... ing body ... ding on ... problem...



1. Intr...

The ... mo...

How mu... to unde... on the s... sion onl... by a few... hours of... this thin... example... derstand... someon...

Pose Embeddings: A Deep Architecture for Learning to Match Human Poses

Greg Mori*[1], Caroline Pantofaru[2], Nisarg Kothari[2], Thomas Leung[2], George Toderici[2], Alexander Toshev[2], and Weilong Yang[2]

[1]Simon Fraser University
[2]Google Inc.

Abstract

We present a method for learning an embedding that places images of humans in similar poses nearby. This embedding can be used as a direct method of comparing im-

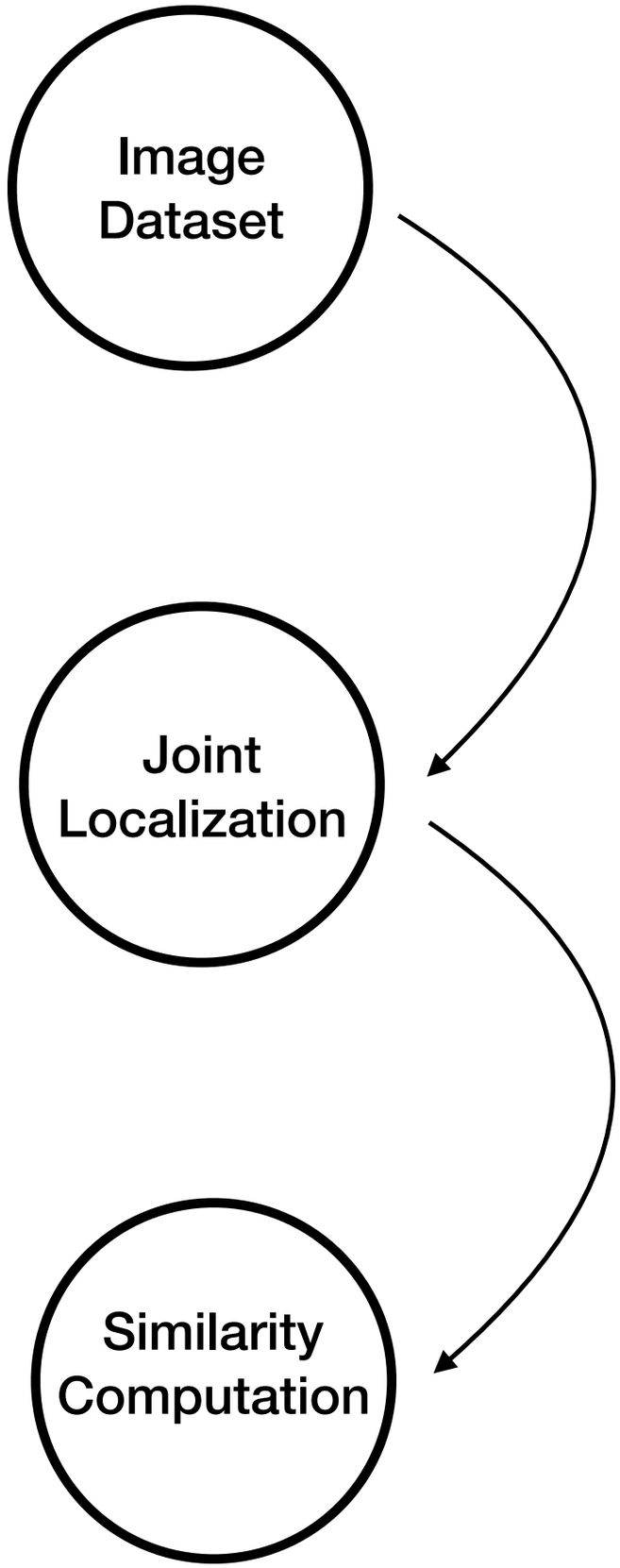


Image
Dataset
Joint
Localization
Similarity
Computation

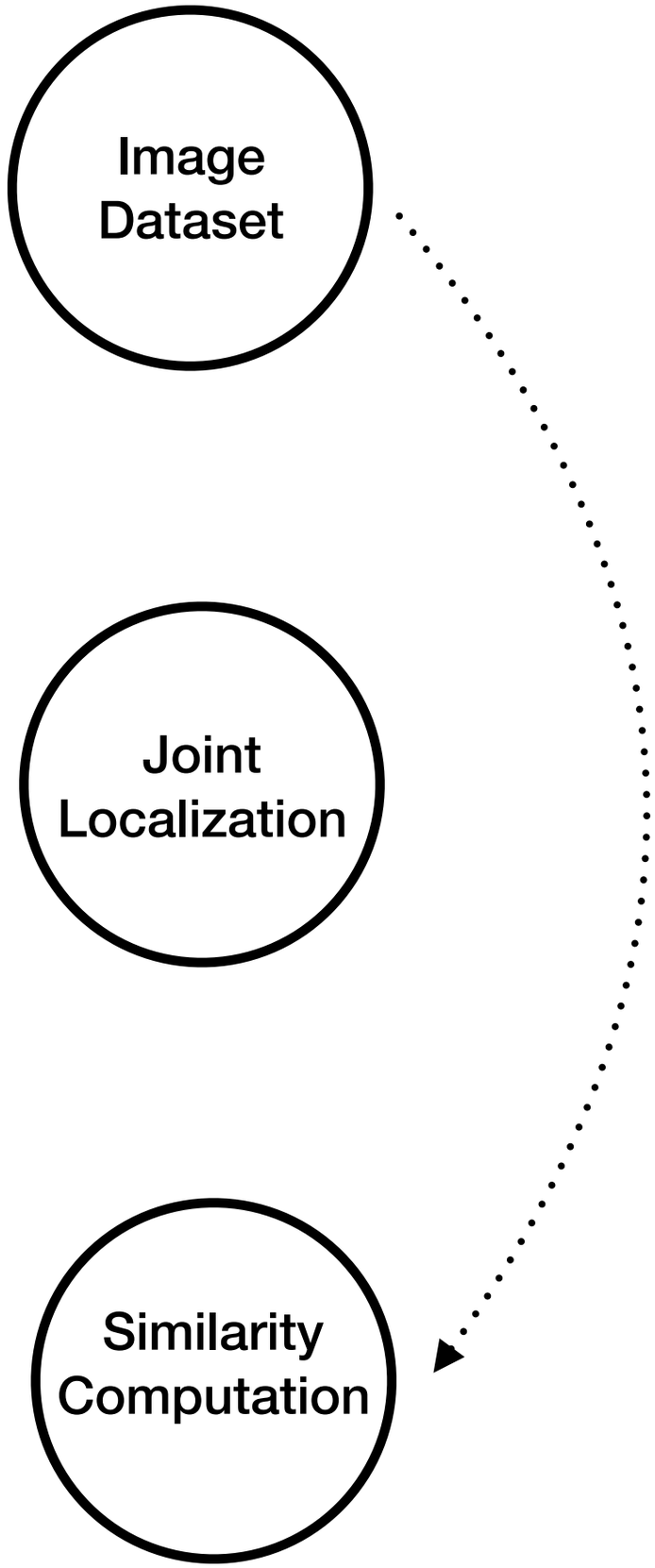
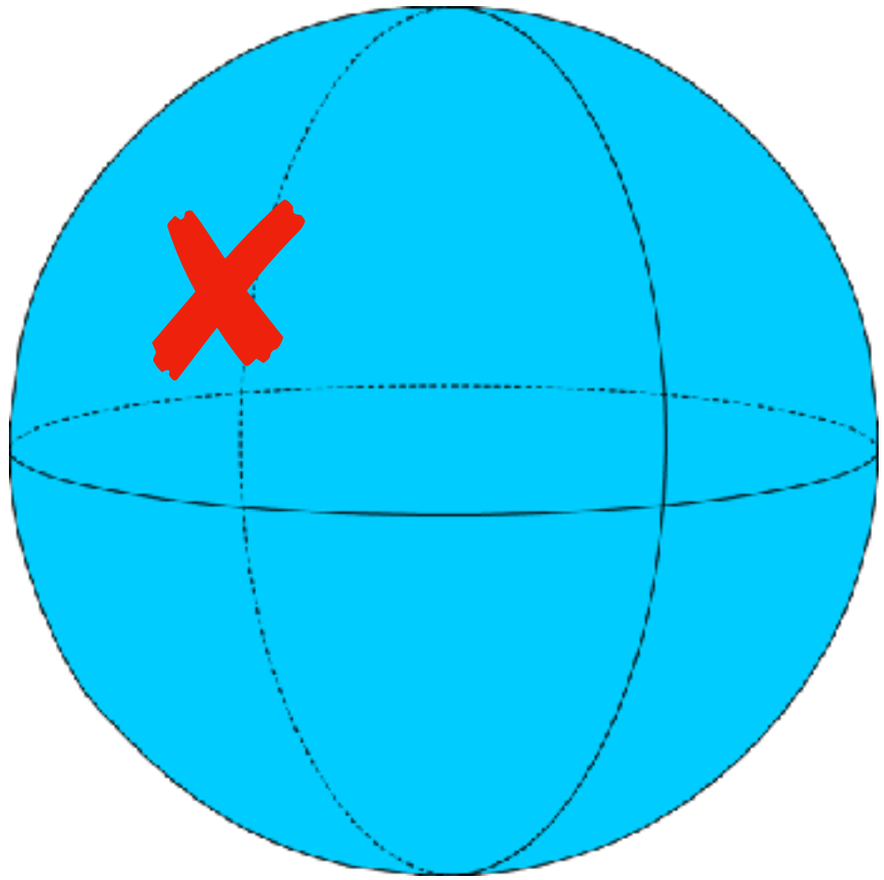
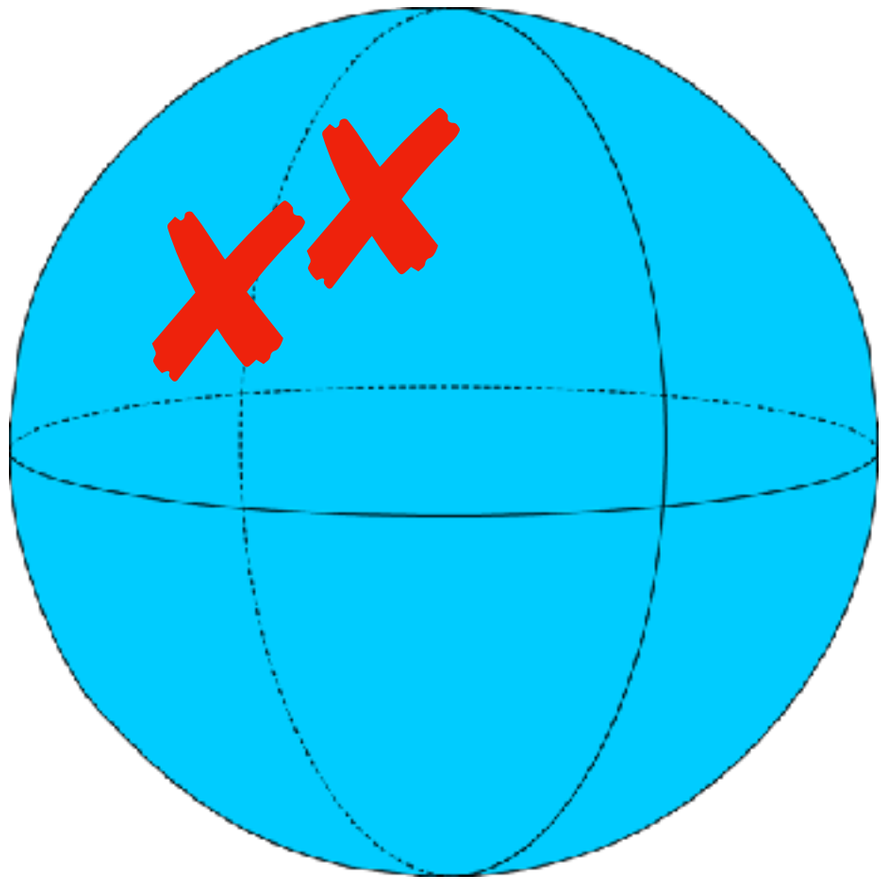
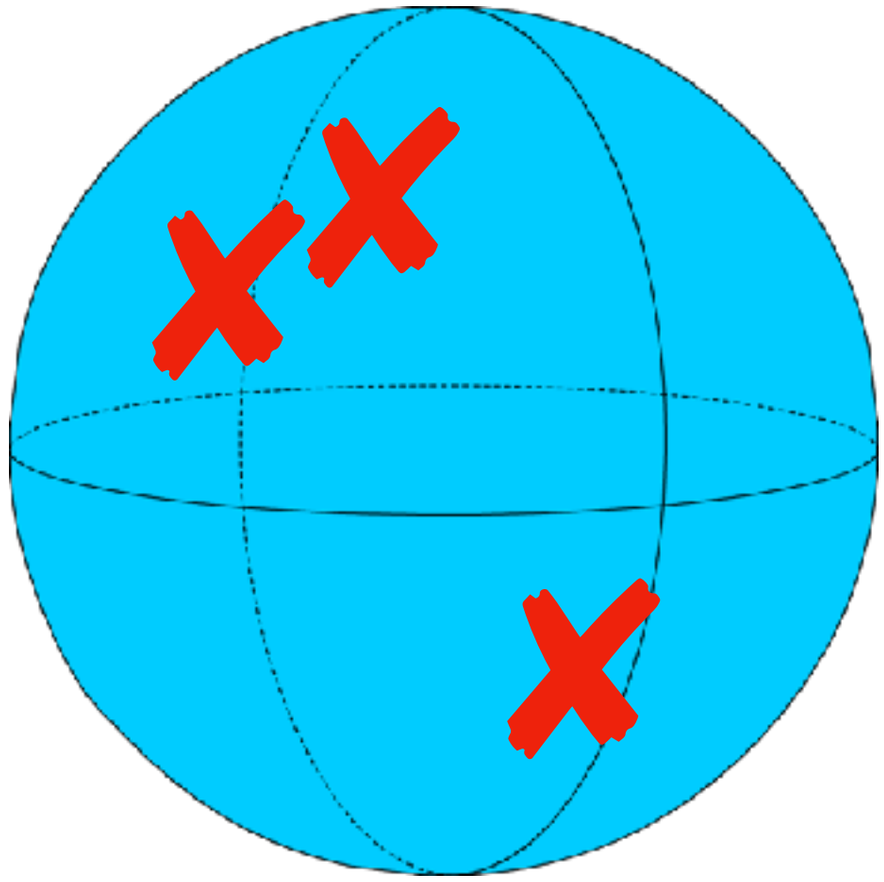
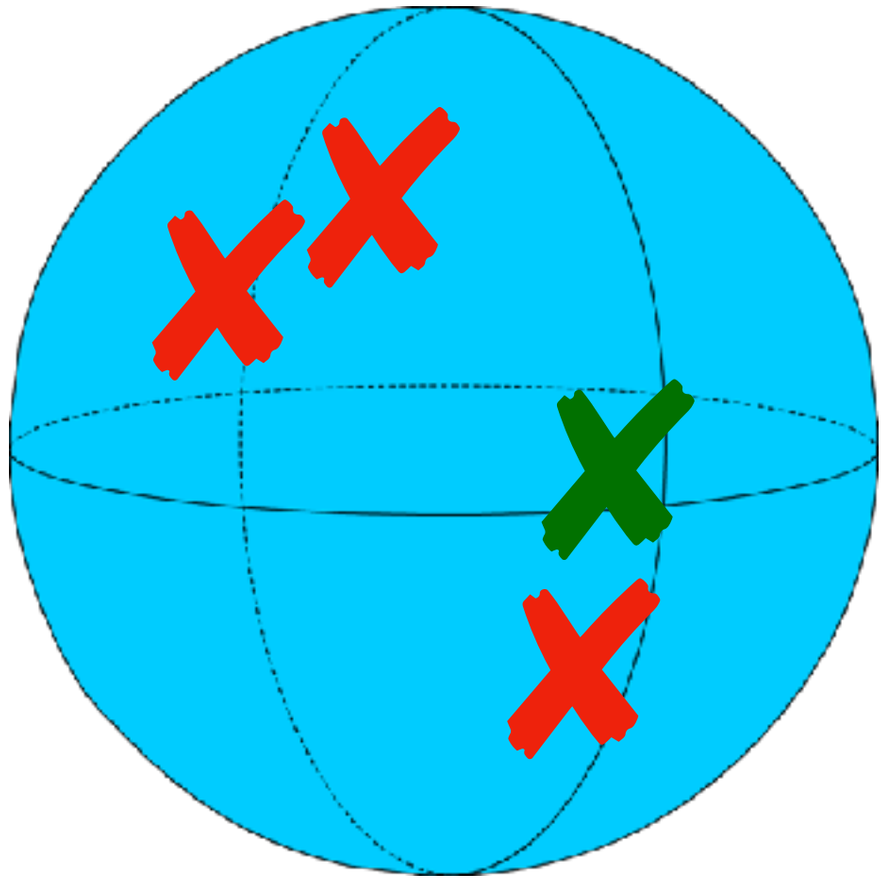


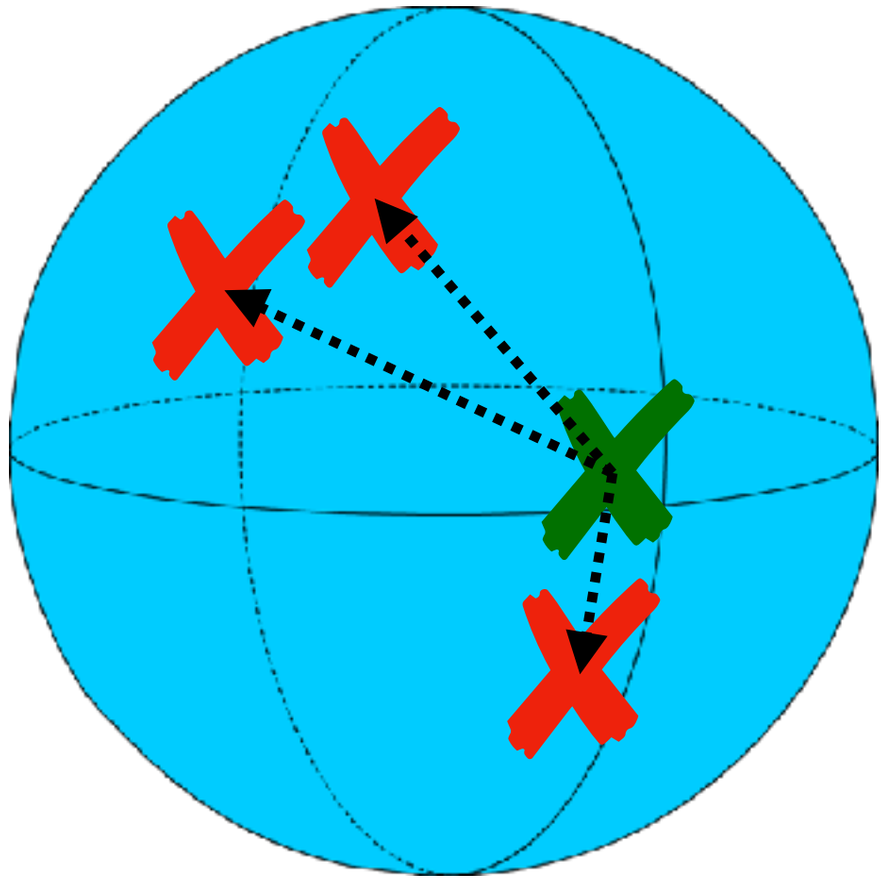
Image Dataset
Joint Localization
Similarity Computation











Main Contributions

Pose-Aware embedding networks

Image-Language embedding network

Two solutions for language subsets

A dataset of 2D, 3D, and Language Primitive descriptors

Background

Baseball

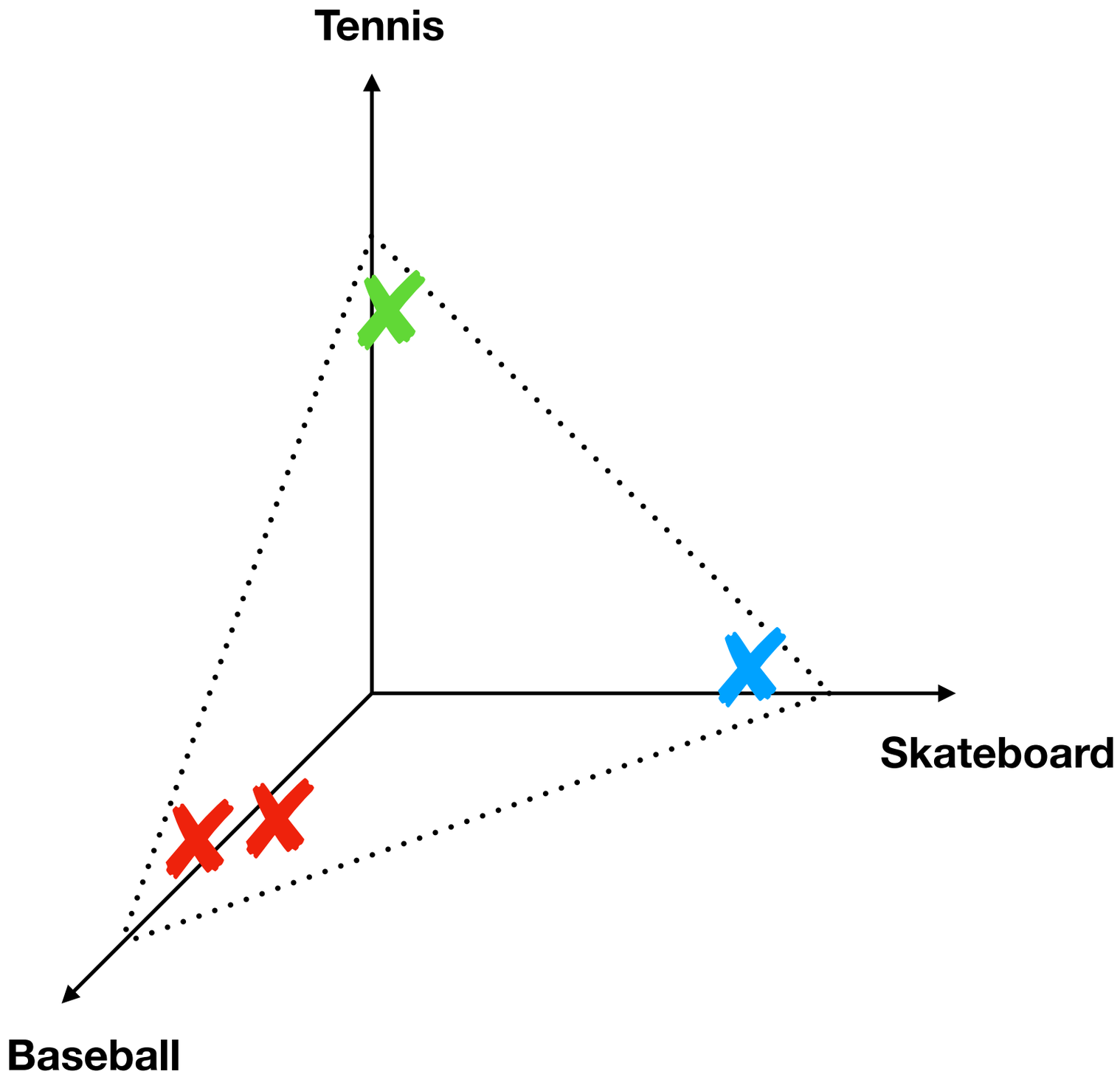
Tennis

Skateboard

0.7 **Baseball**

0.2 **Tennis**

0.1 **Skateboard**

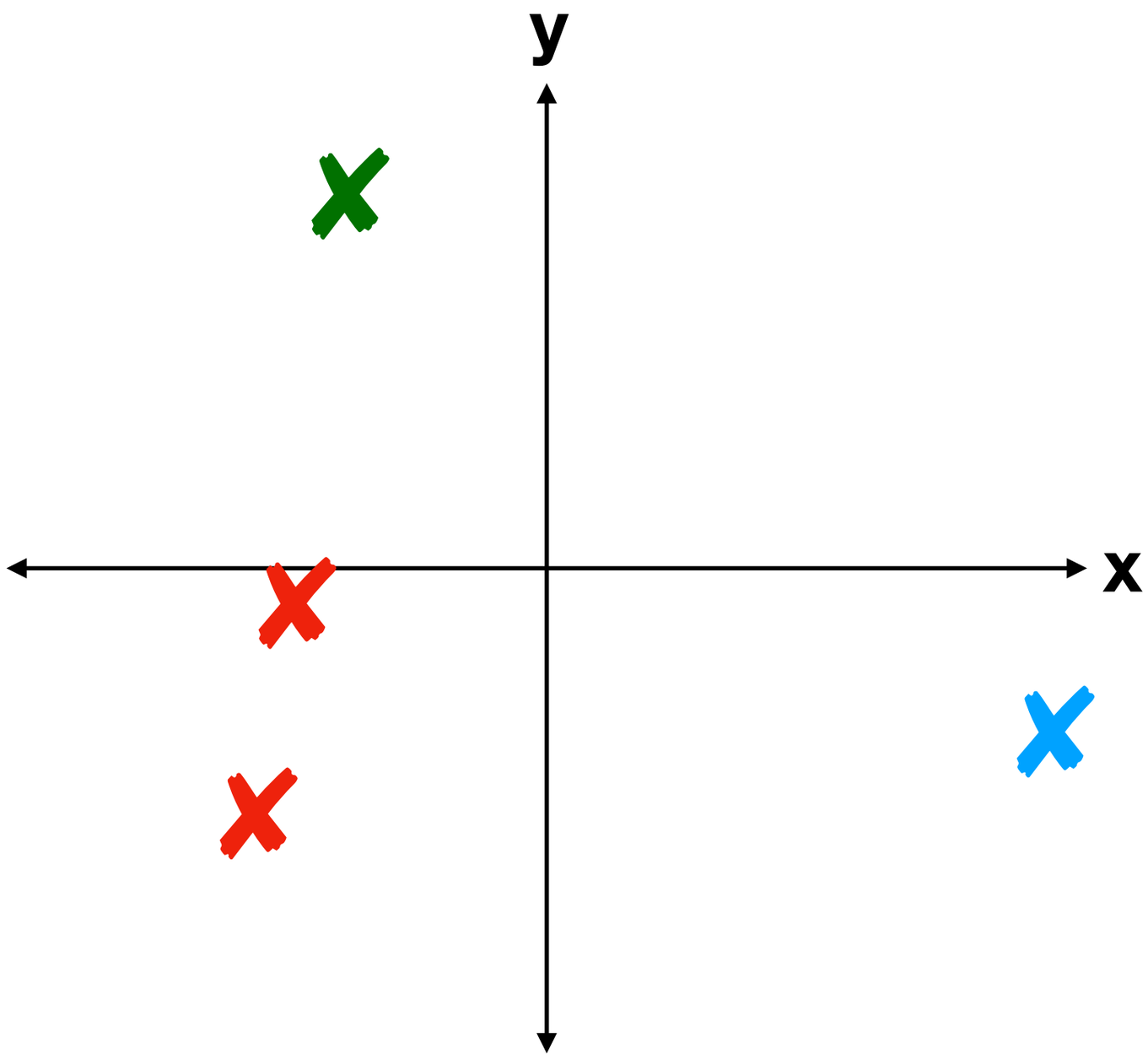

Tennis
Skateboard
Baseball

x

y

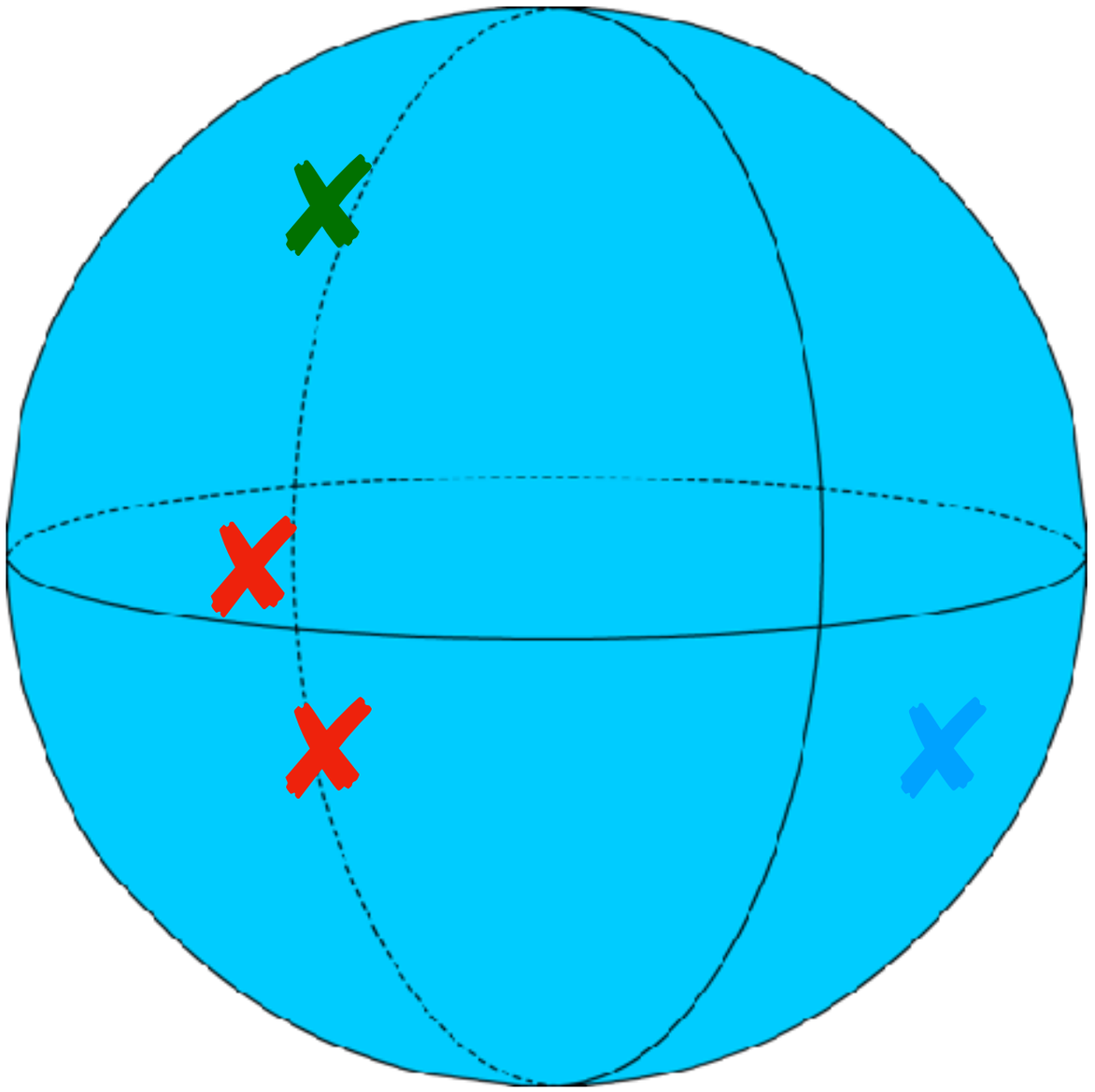
0.3 x

0.9 y



$$||f(a) - f(p)||_2^2 < ||f(a) - f(n)||_2^2$$

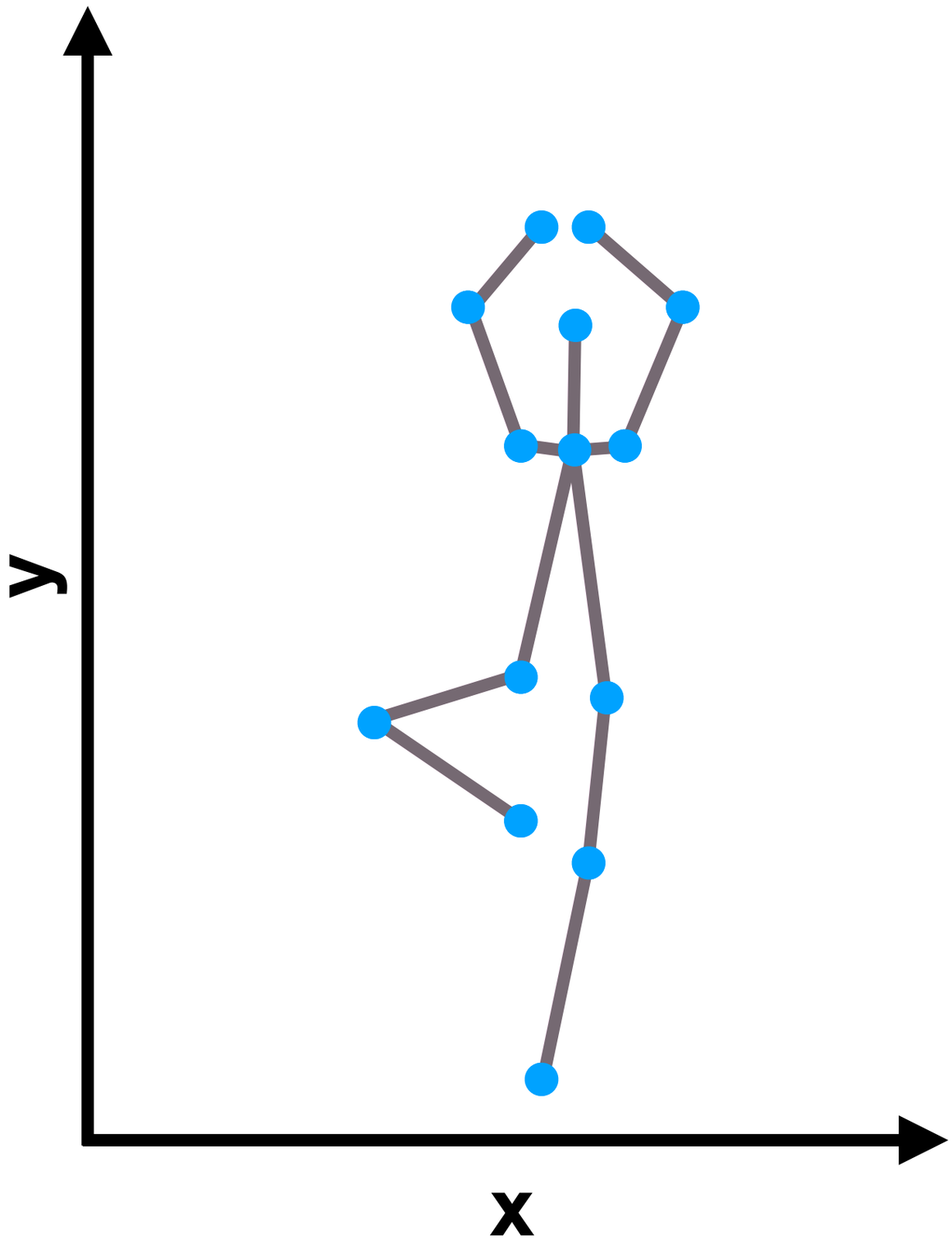
$$\forall a, p, n \in D \; s.t. \; \delta(a_{pose}, p_{pose}) < \delta(a_{pose}, n_{pose})$$

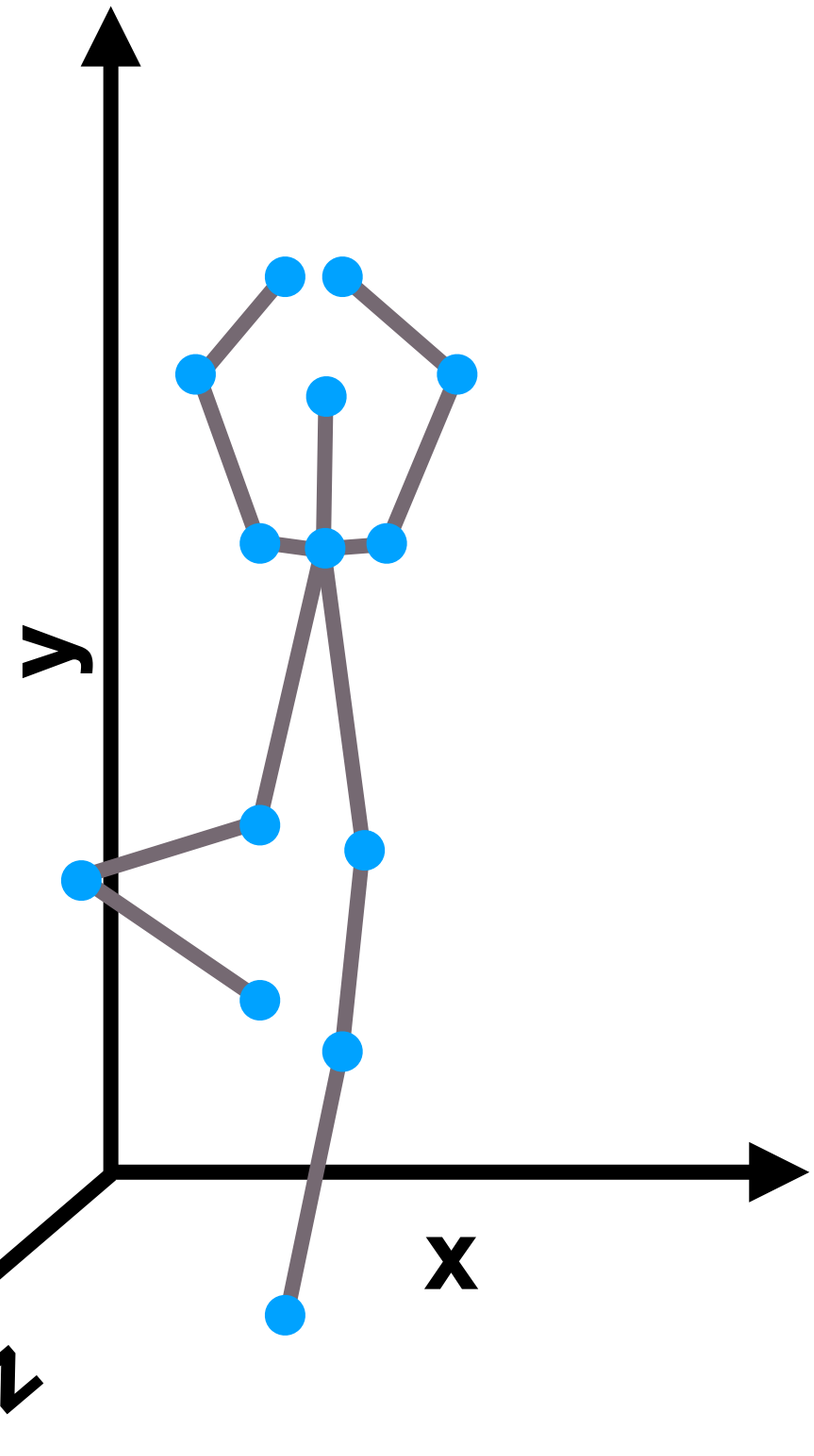


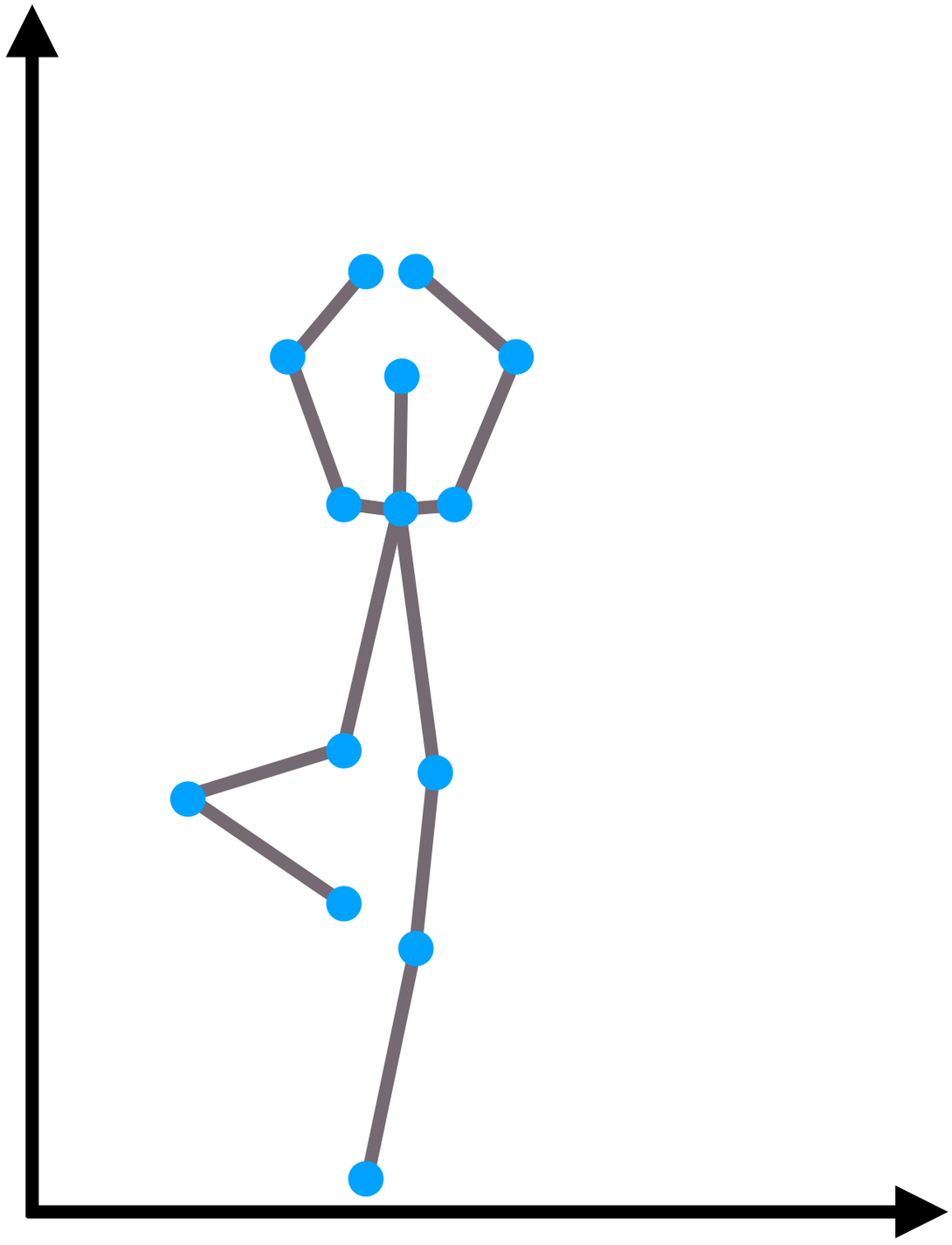
$$||f(a) - f(p)||_2^2 < ||f(a) - f(n)||_2^2$$

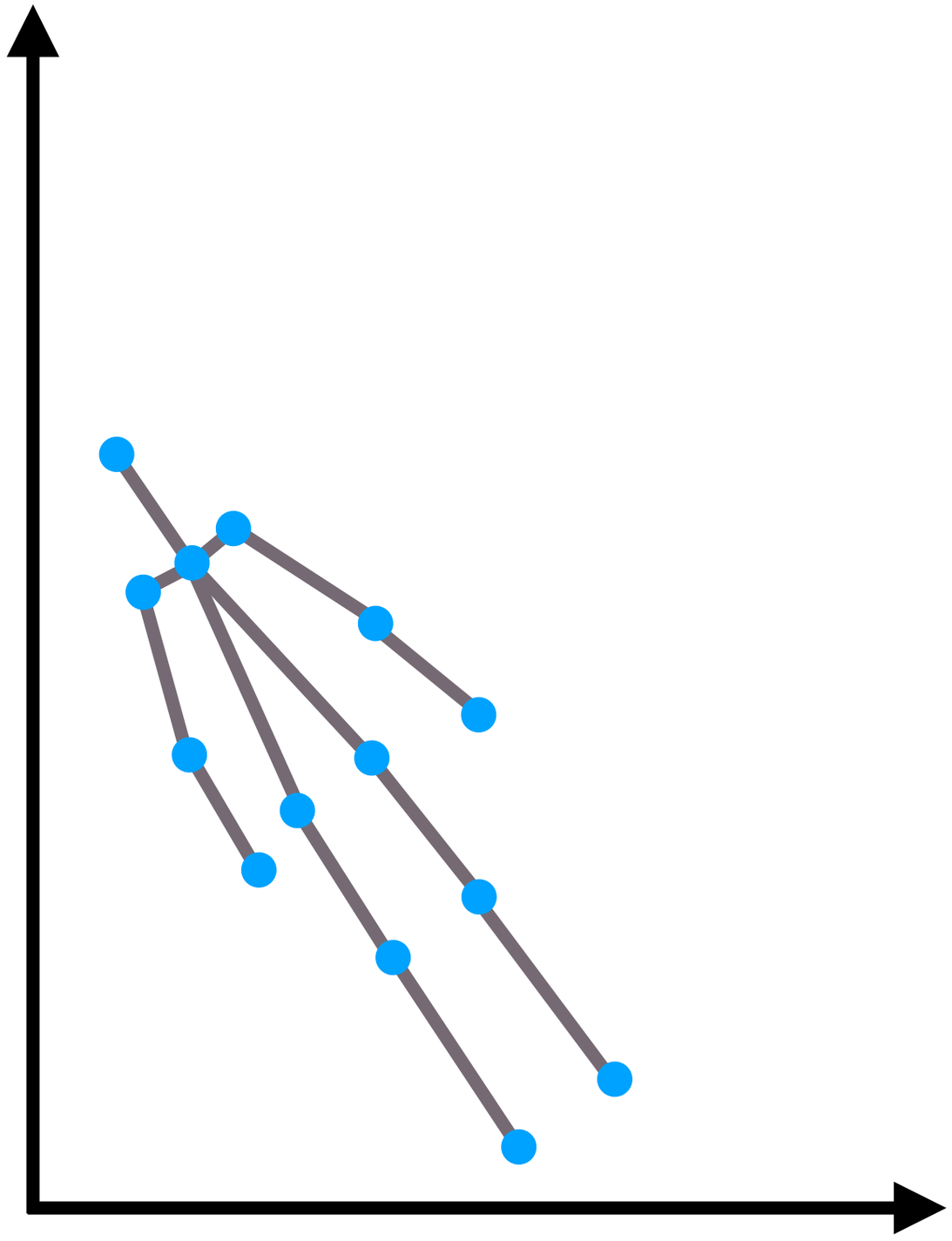
$$\forall a, p, n \ \in D \ s.t. \ \delta(a_{pose}, p_{pose}) < \delta(a_{pose}, n_{pose})$$

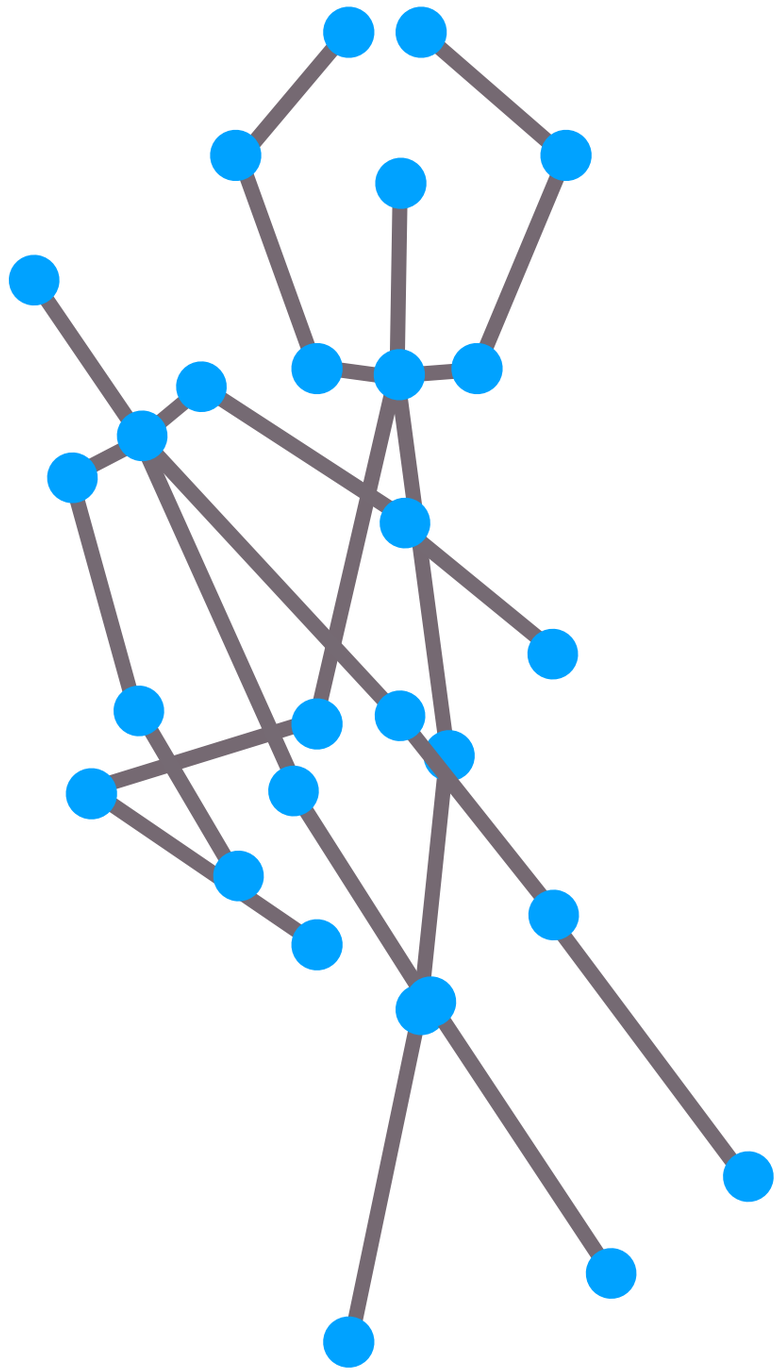
Image Querying


y
x

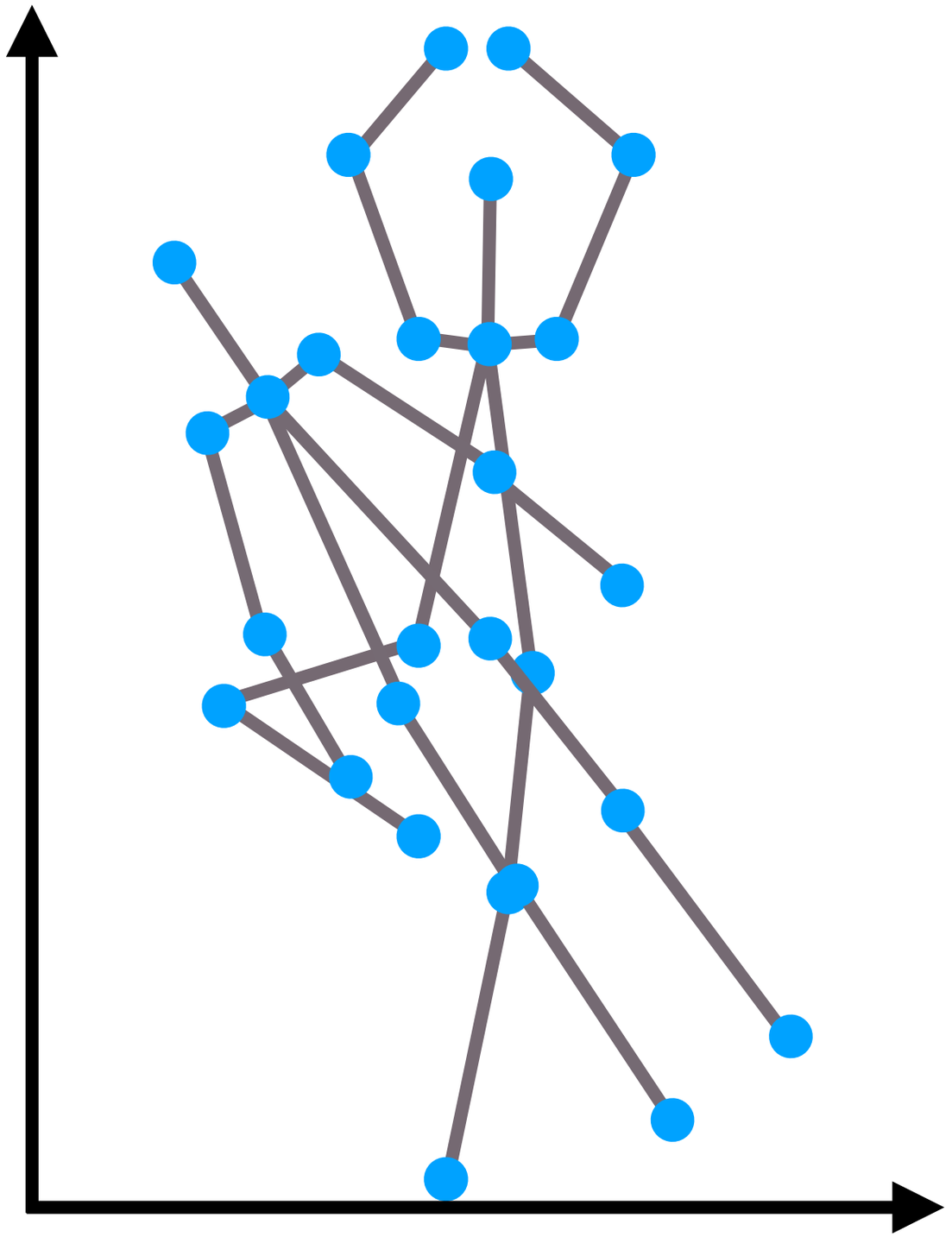

y
x
z



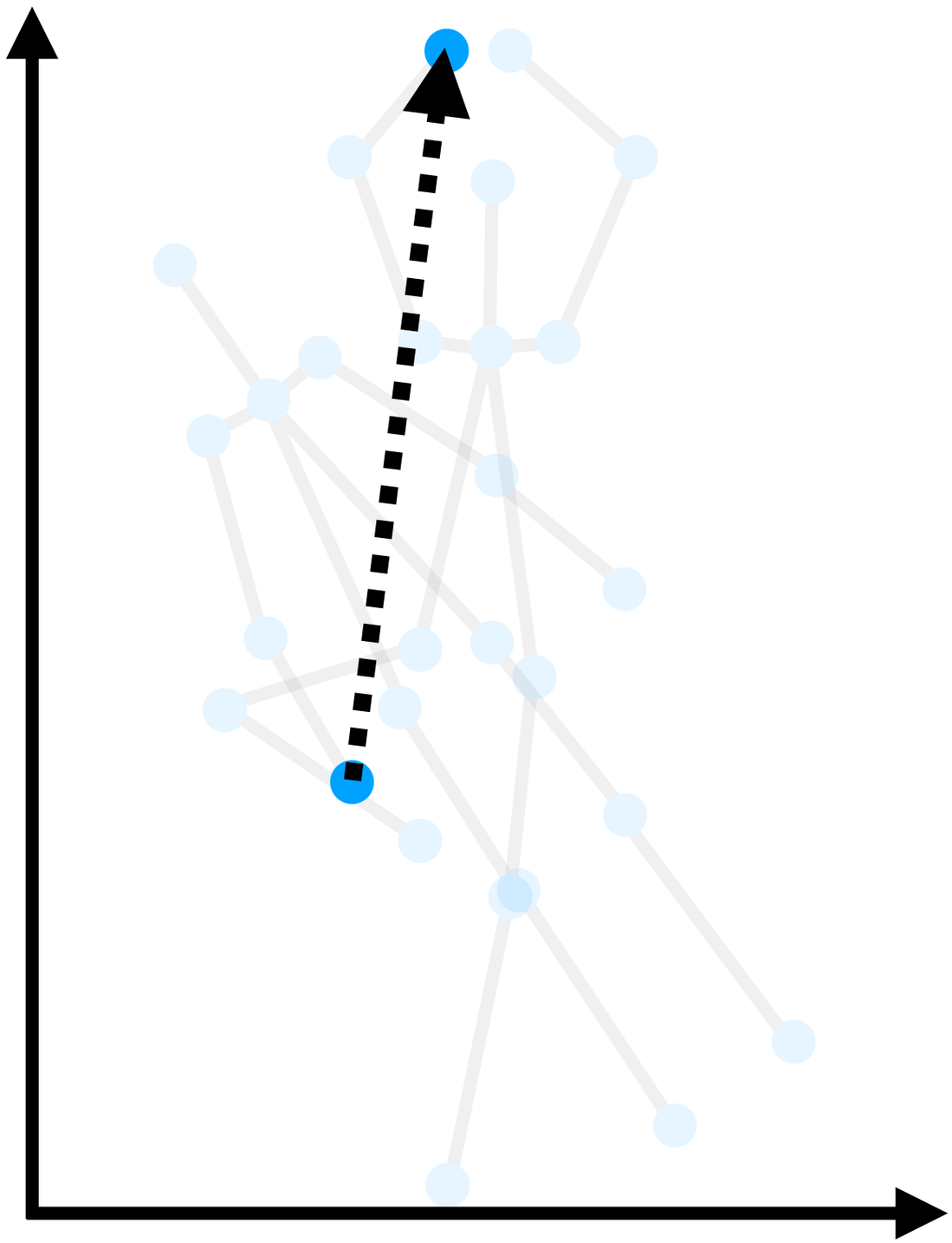




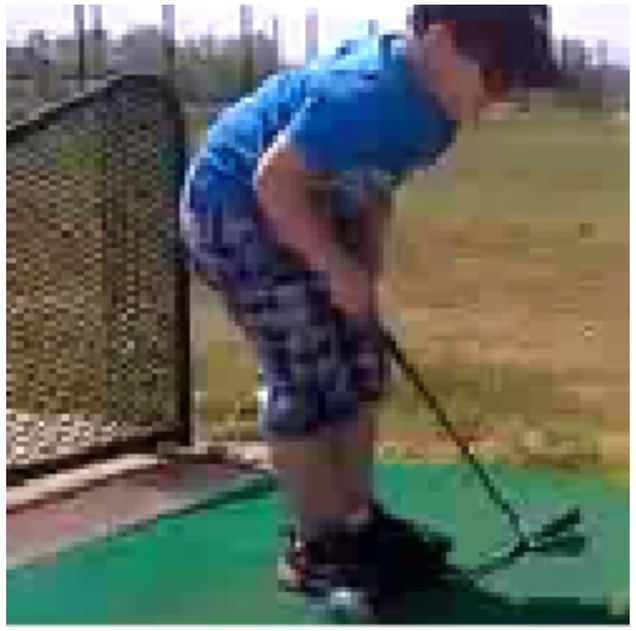
$$\delta_{2D}(J_a, J_b) = \frac{1}{N} \sum_{i=1}^{N} \sqrt{|J_a^i - J_b^i|^2}$$

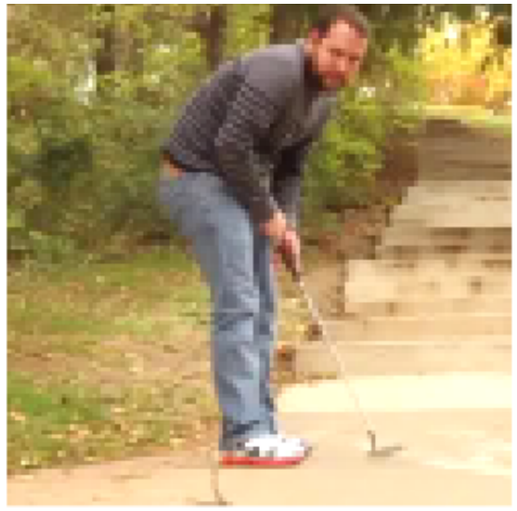
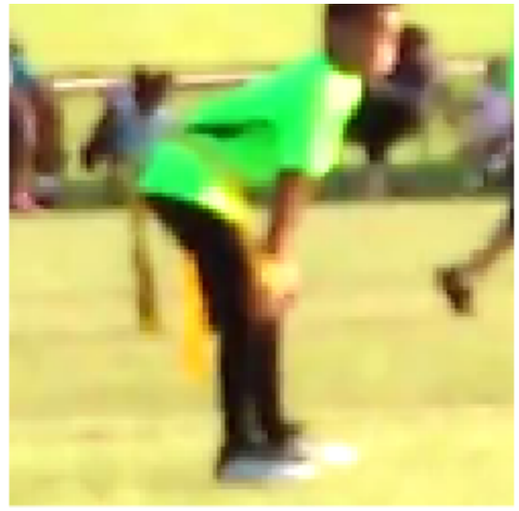
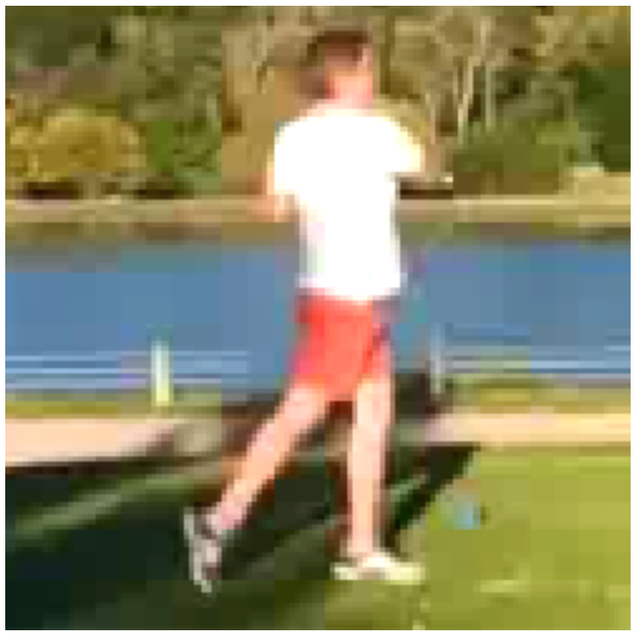


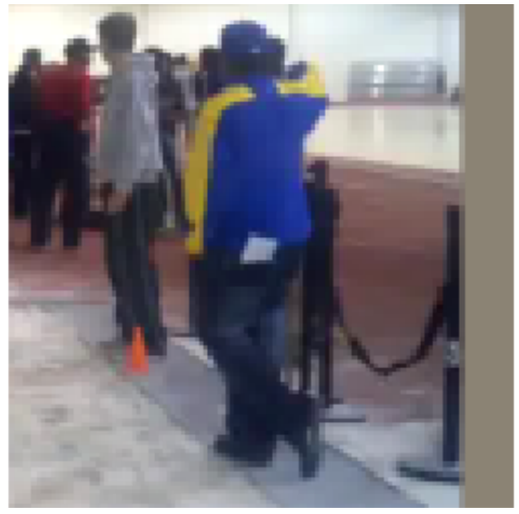
$$\delta_{2D}(J_a, J_b) = \frac{1}{N}\sum_{i=1}^{N}\sqrt{|J_a^i - J_b^i|^2}$$



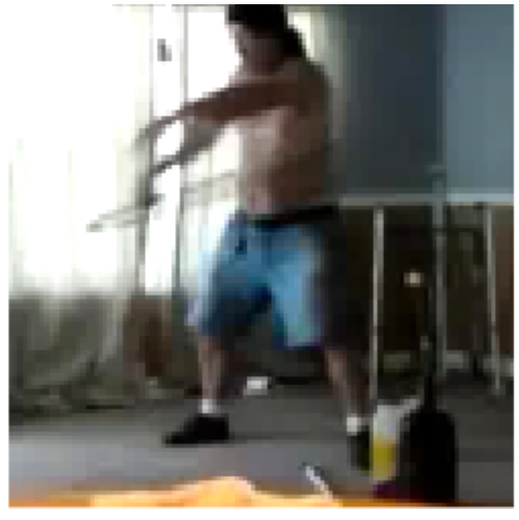
$$\delta_{2D}(J_a, J_b) = \frac{1}{N}\sum_{i=1}^{N}\sqrt{|J_a^i - J_b^i|^2}$$

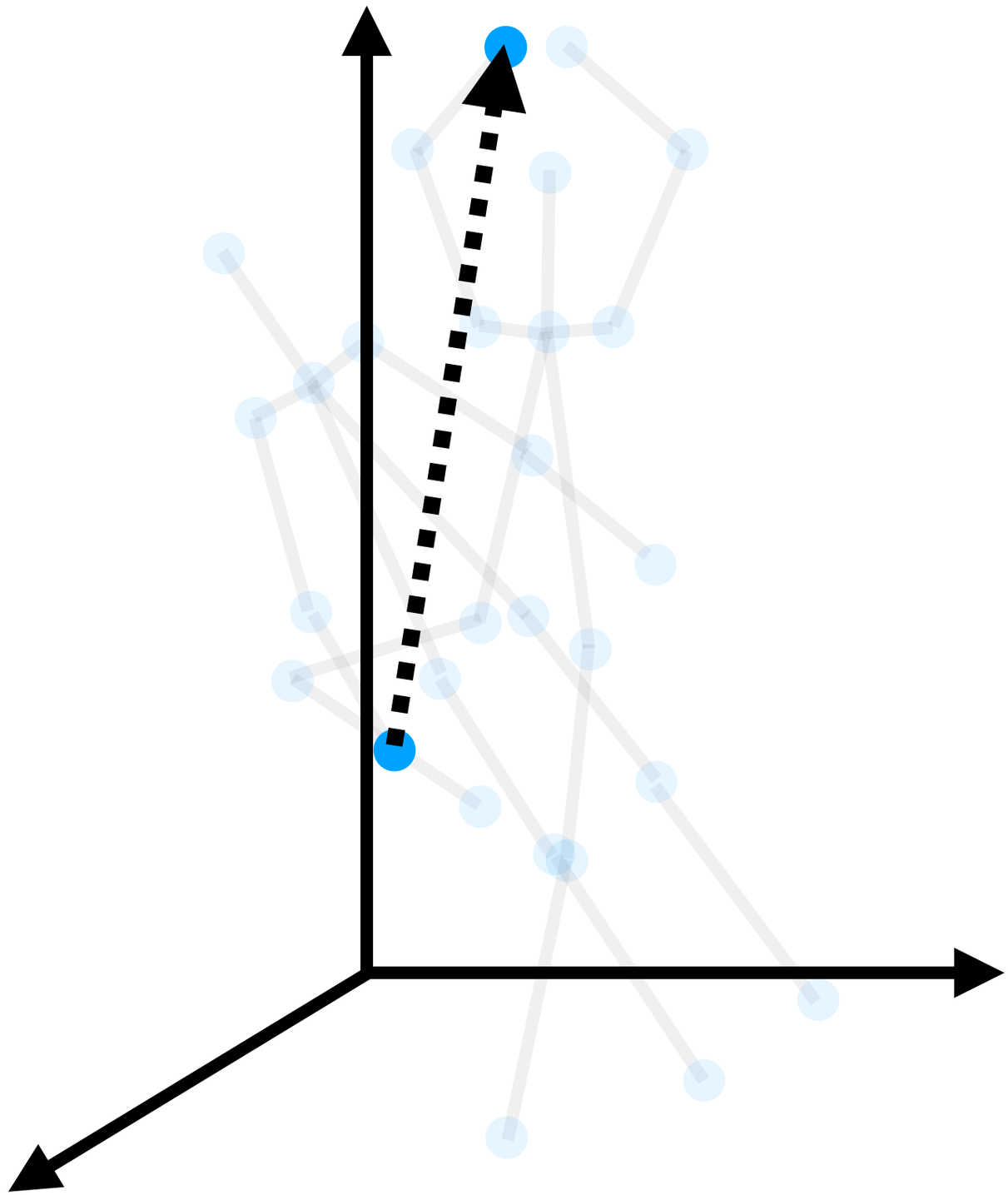


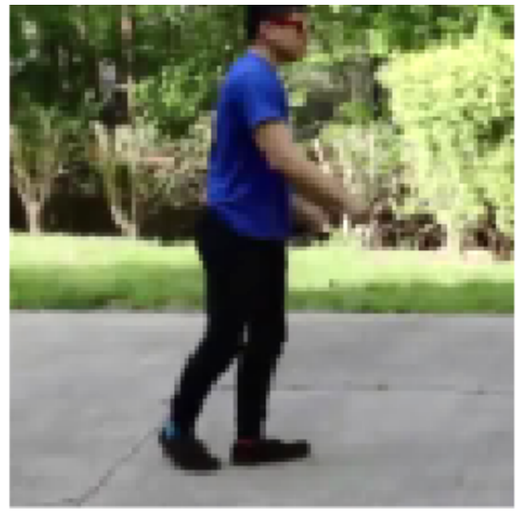
www.wildfilmsinc

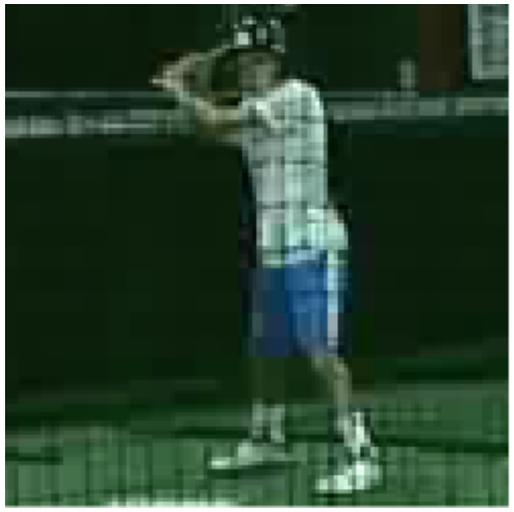
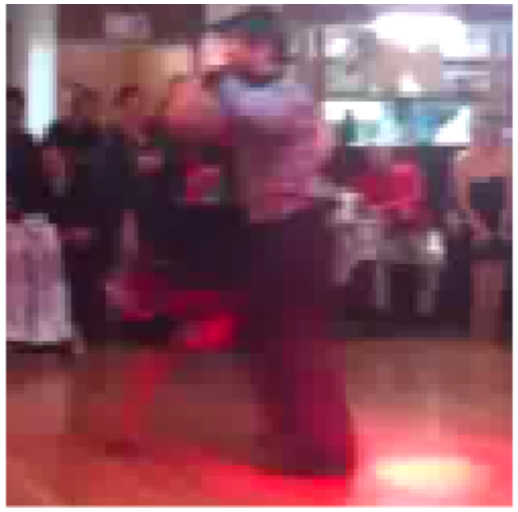
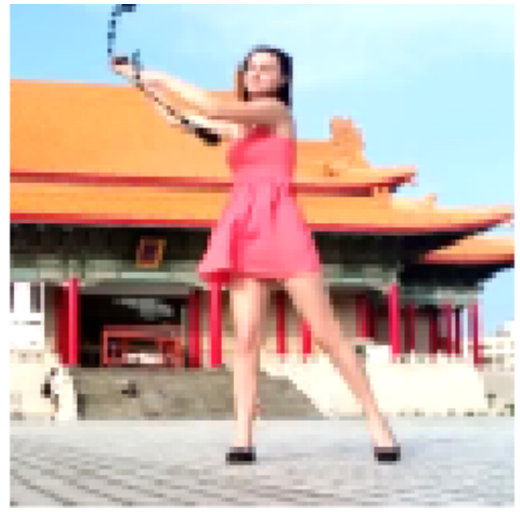
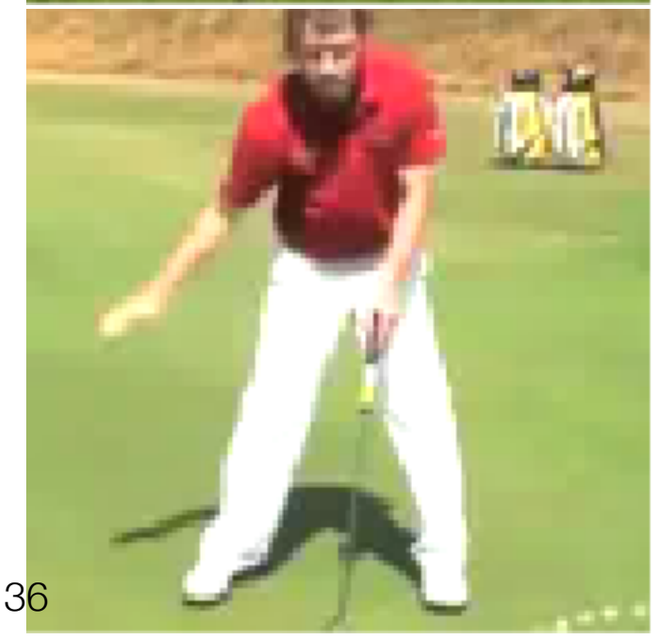
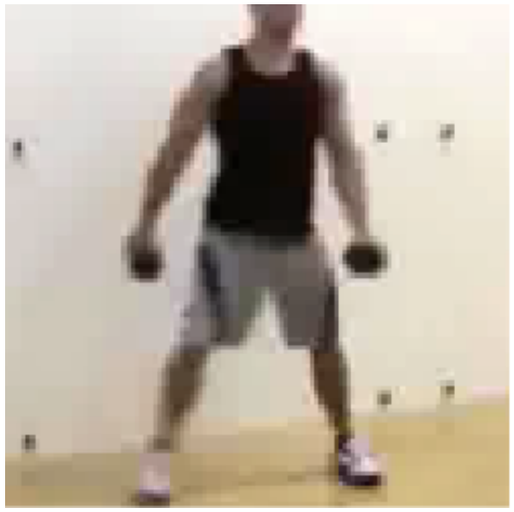
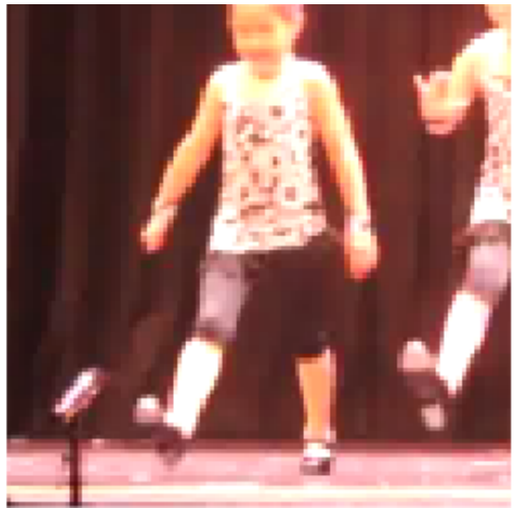
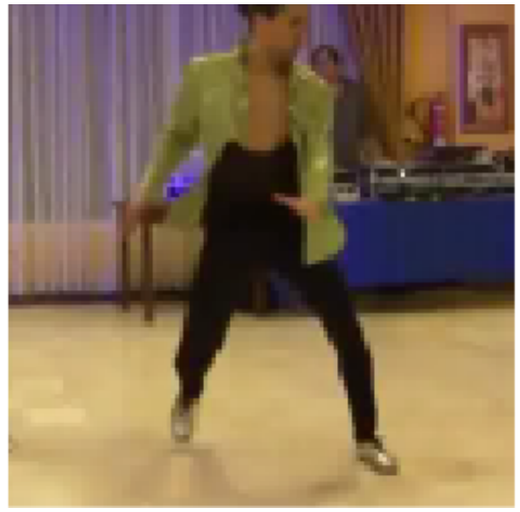



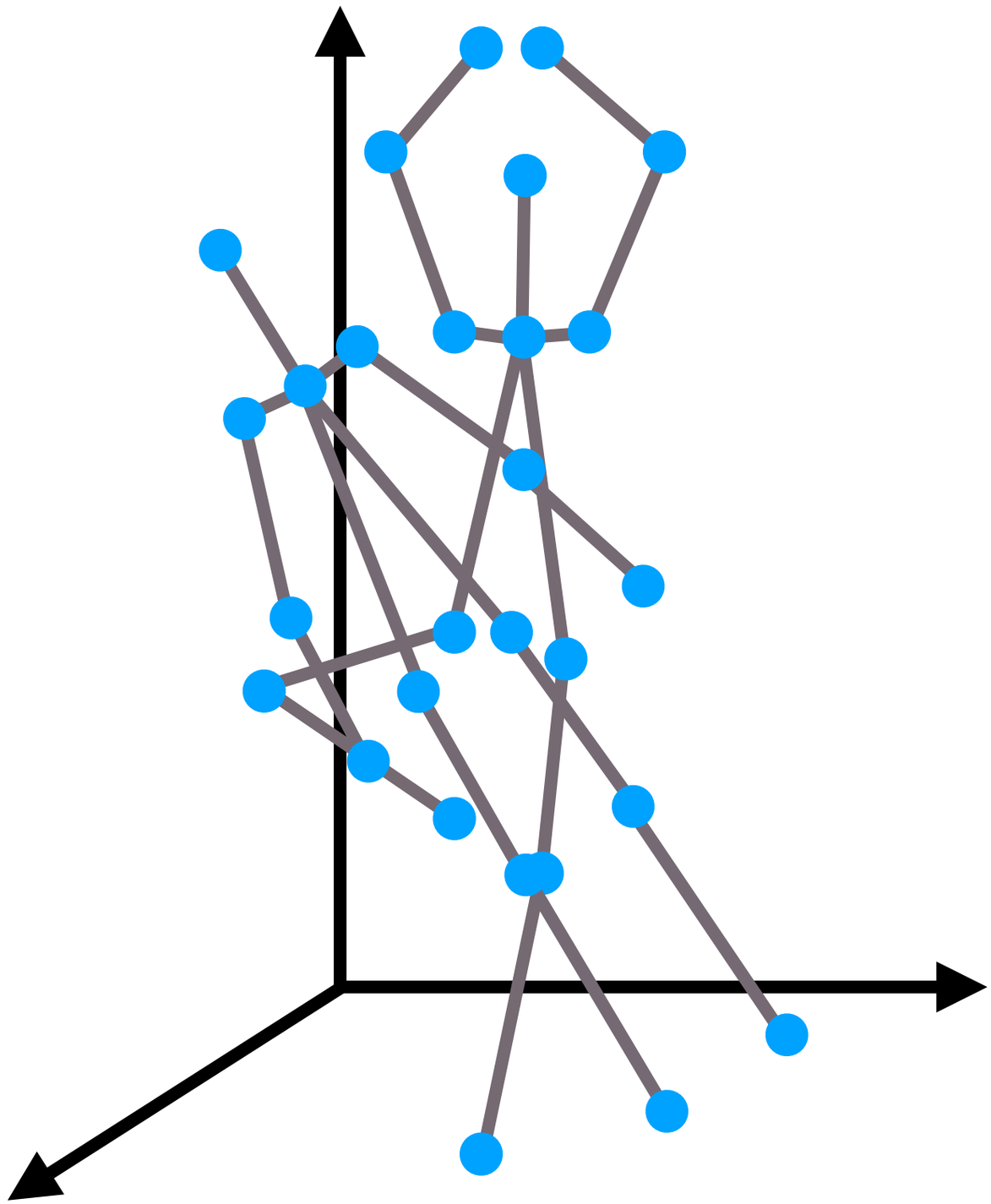


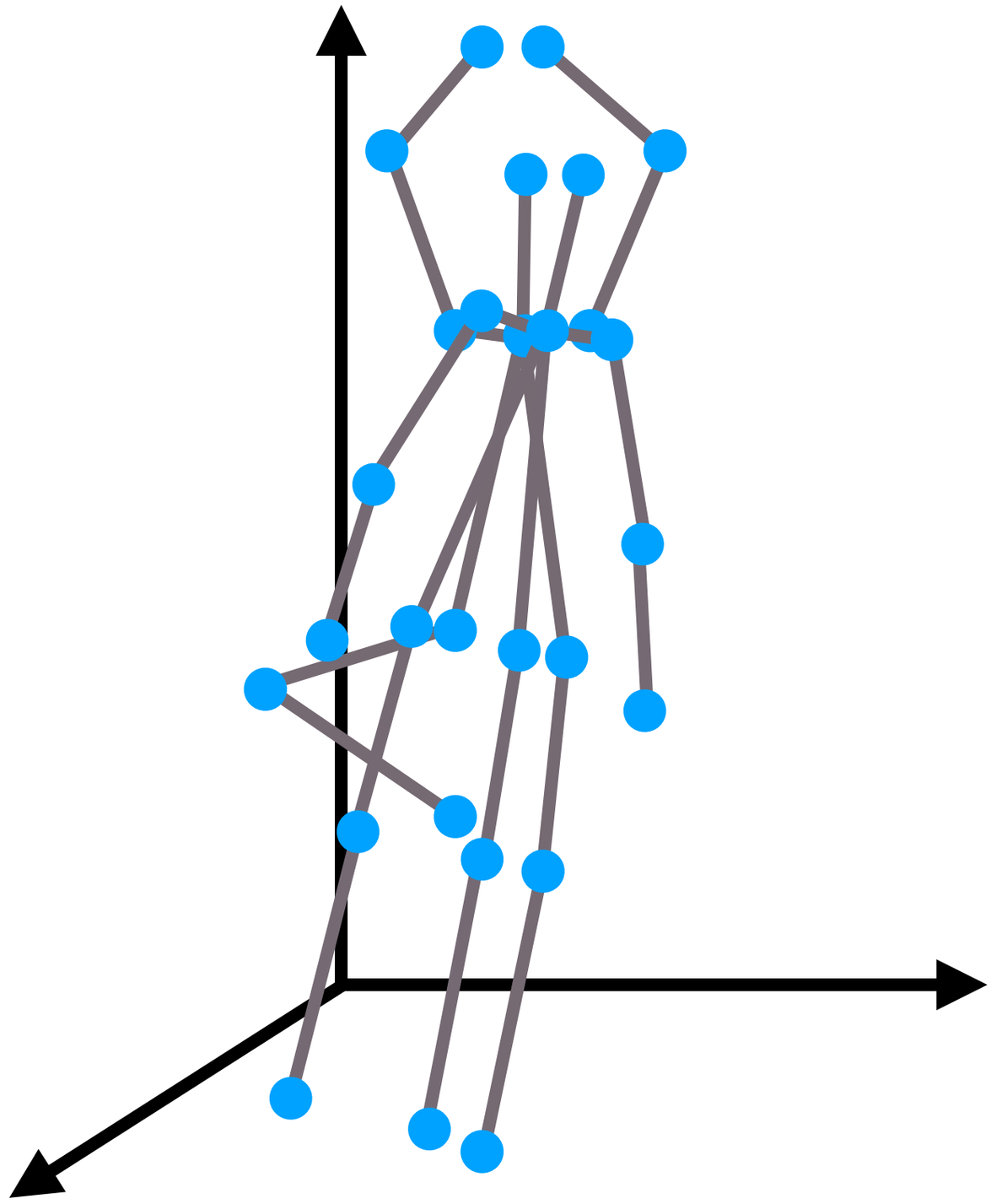
$$\delta_{3D}(J_a, J_b) = \frac{1}{N}\sum_{i=1}^{N}\sqrt{|J_a^i - J_b^i|^2}$$



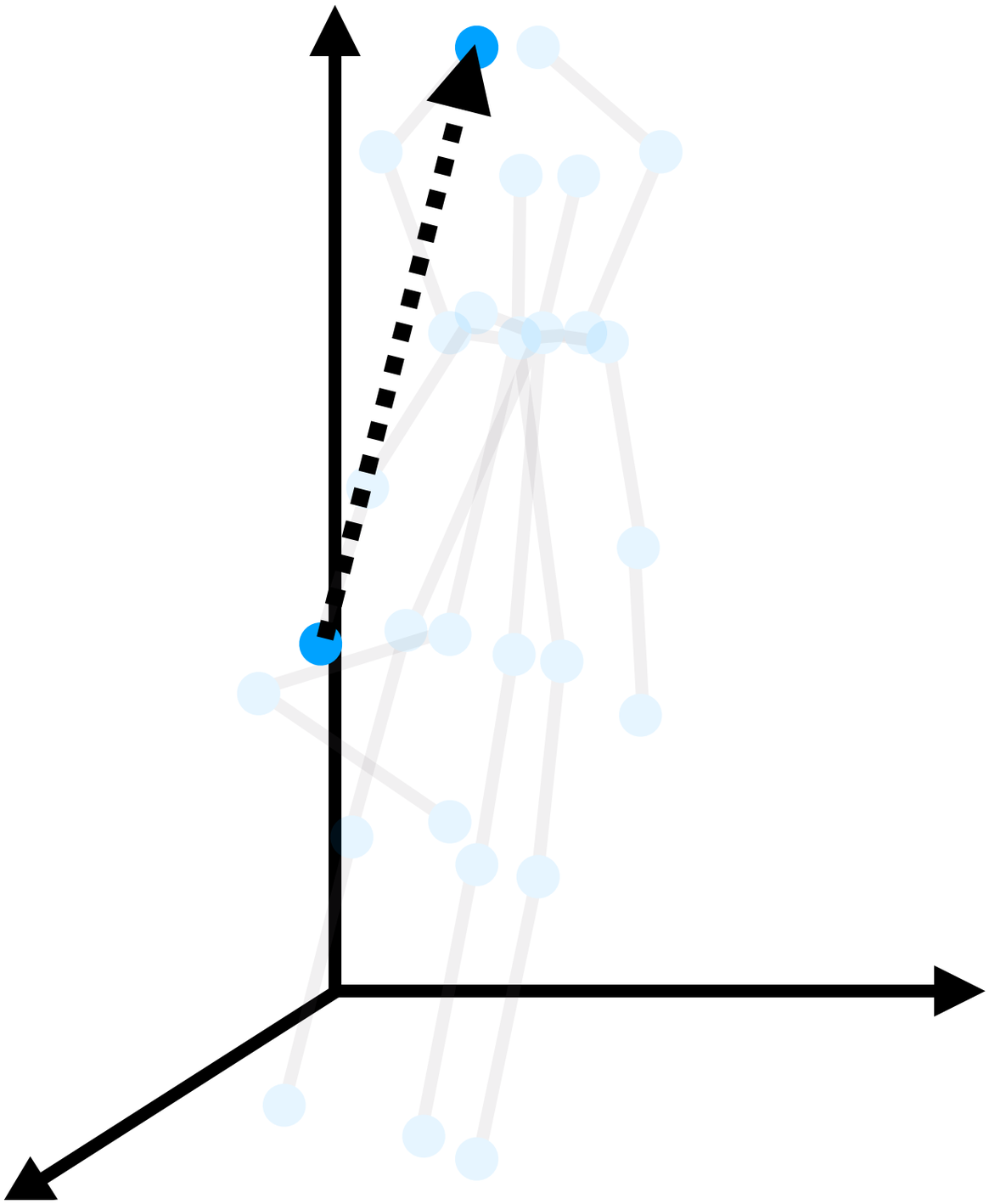



www.wildfilmsindi








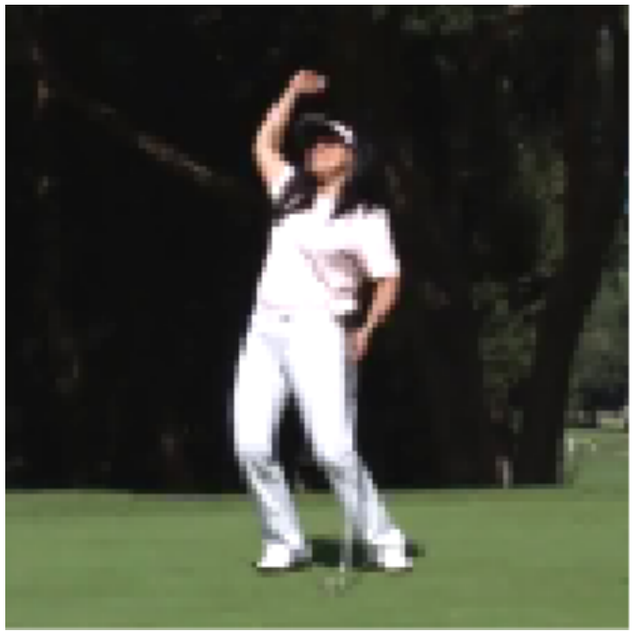
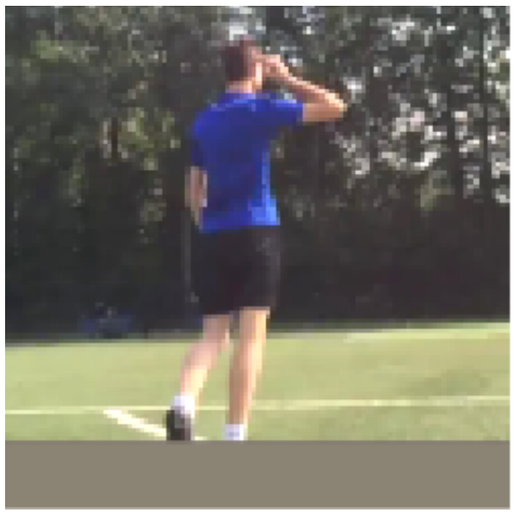


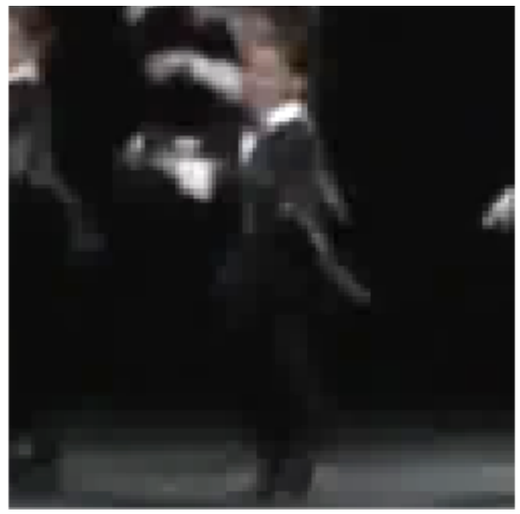


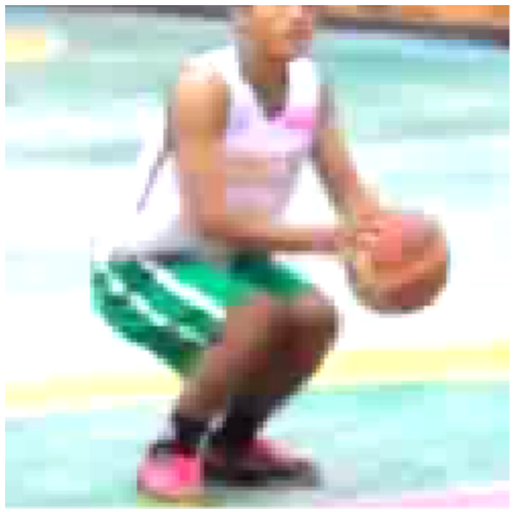
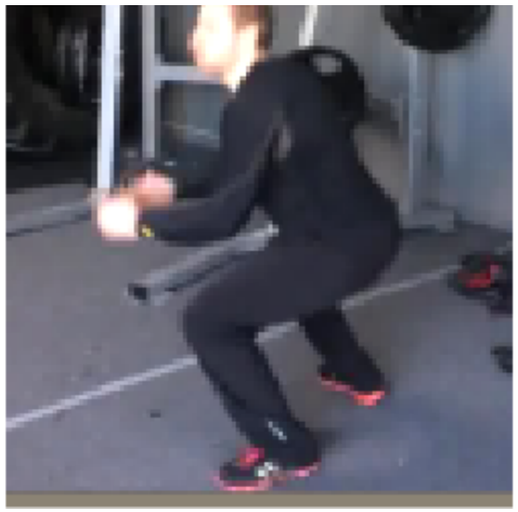
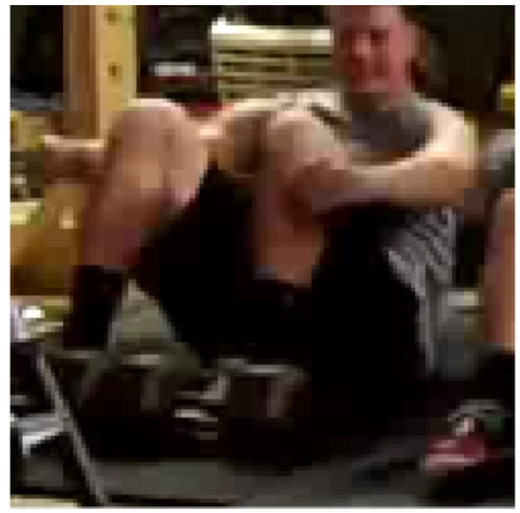
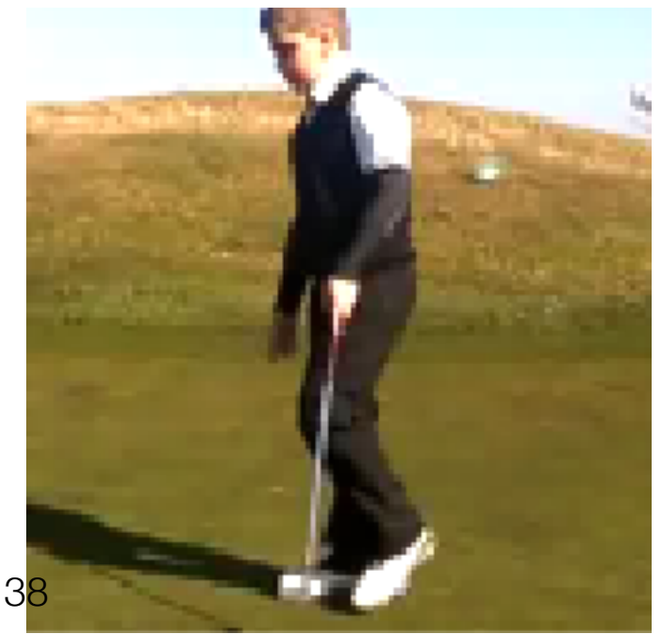
$$P(J_b^i; J_a^i)$$

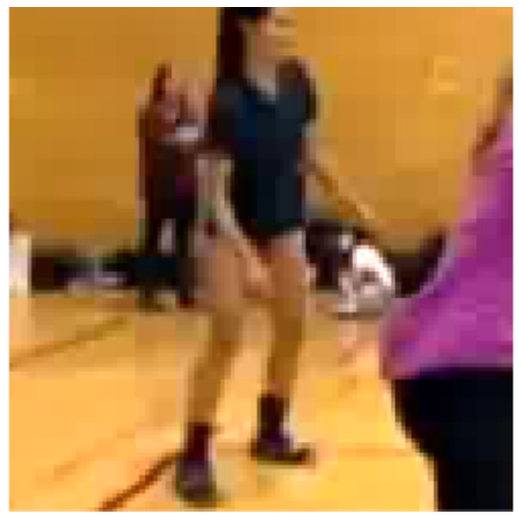


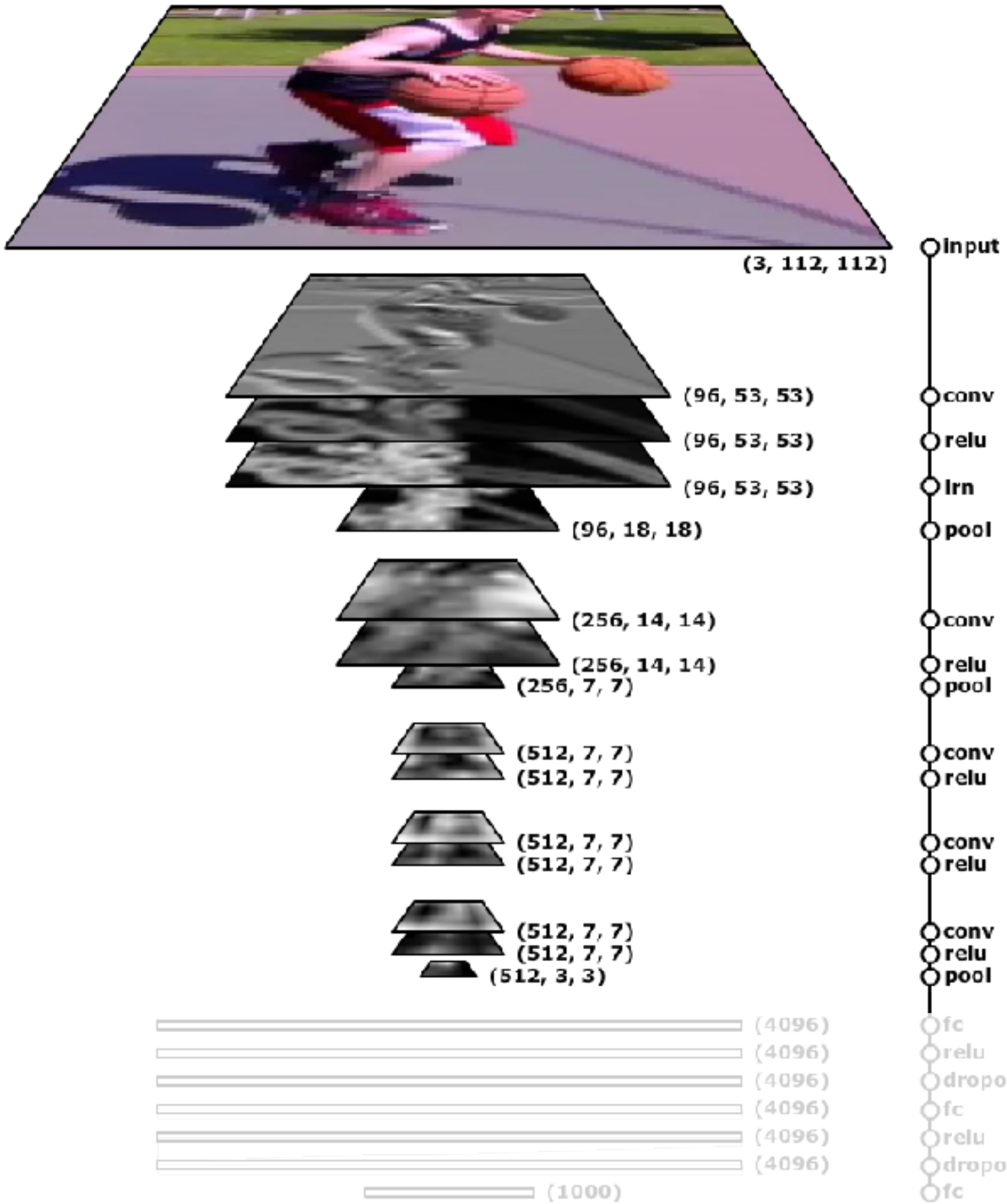
$$\delta_{procrustes}(J_a, J_b) = \sum_{i=1}^{N} \sqrt{|J_a^i - P(J_b^i; J_a^i)|^2}$$

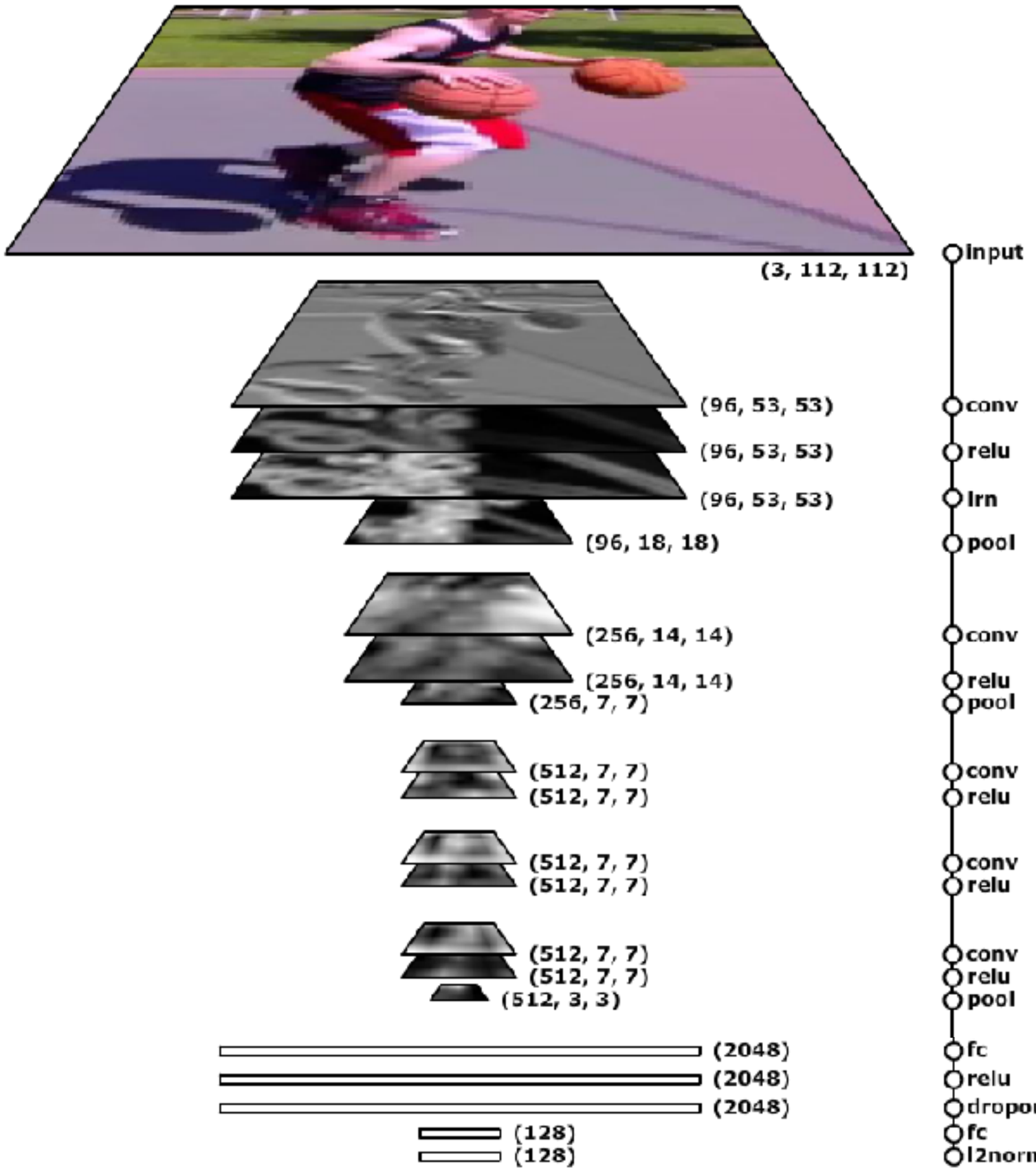





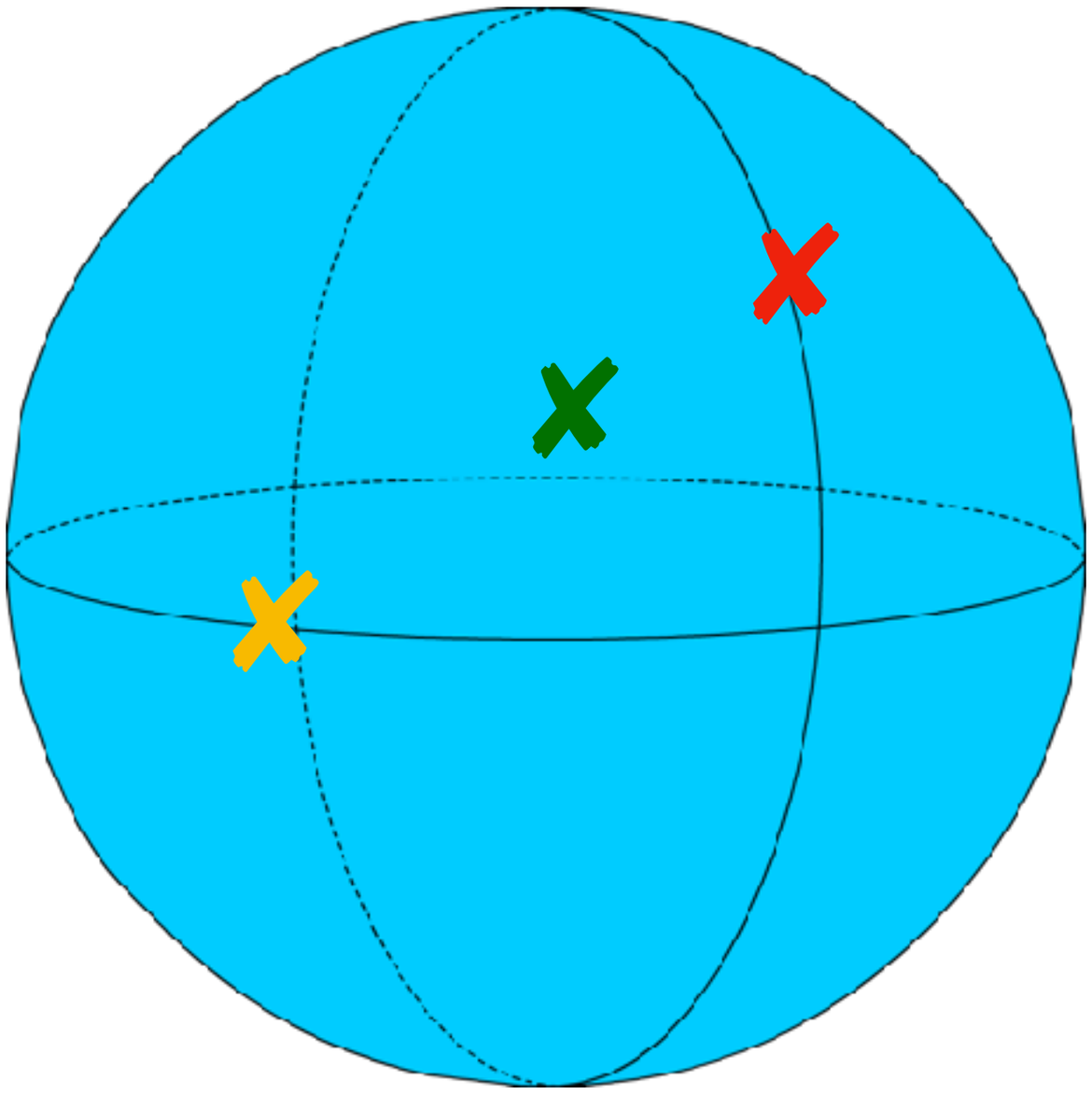




Input
(3, 112, 112)
conv
(96, 53, 53)
relu
(96, 53, 53)
lrn
(96, 53, 53)
pool
(96, 18, 18)
conv
(256, 14, 14)
relu
(256, 14, 14)
pool
(256, 7, 7)
conv
(512, 7, 7)
relu
(512, 7, 7)
conv
(512, 7, 7)
relu
(512, 7, 7)
conv
(512, 7, 7)
relu
(512, 7, 7)
pool
(512, 3, 3)
fc
(4096)
relu
(4096)
dropout
(4096)
fc
(4096)
relu
(4096)
dropout
(4096)
fc
(1000)

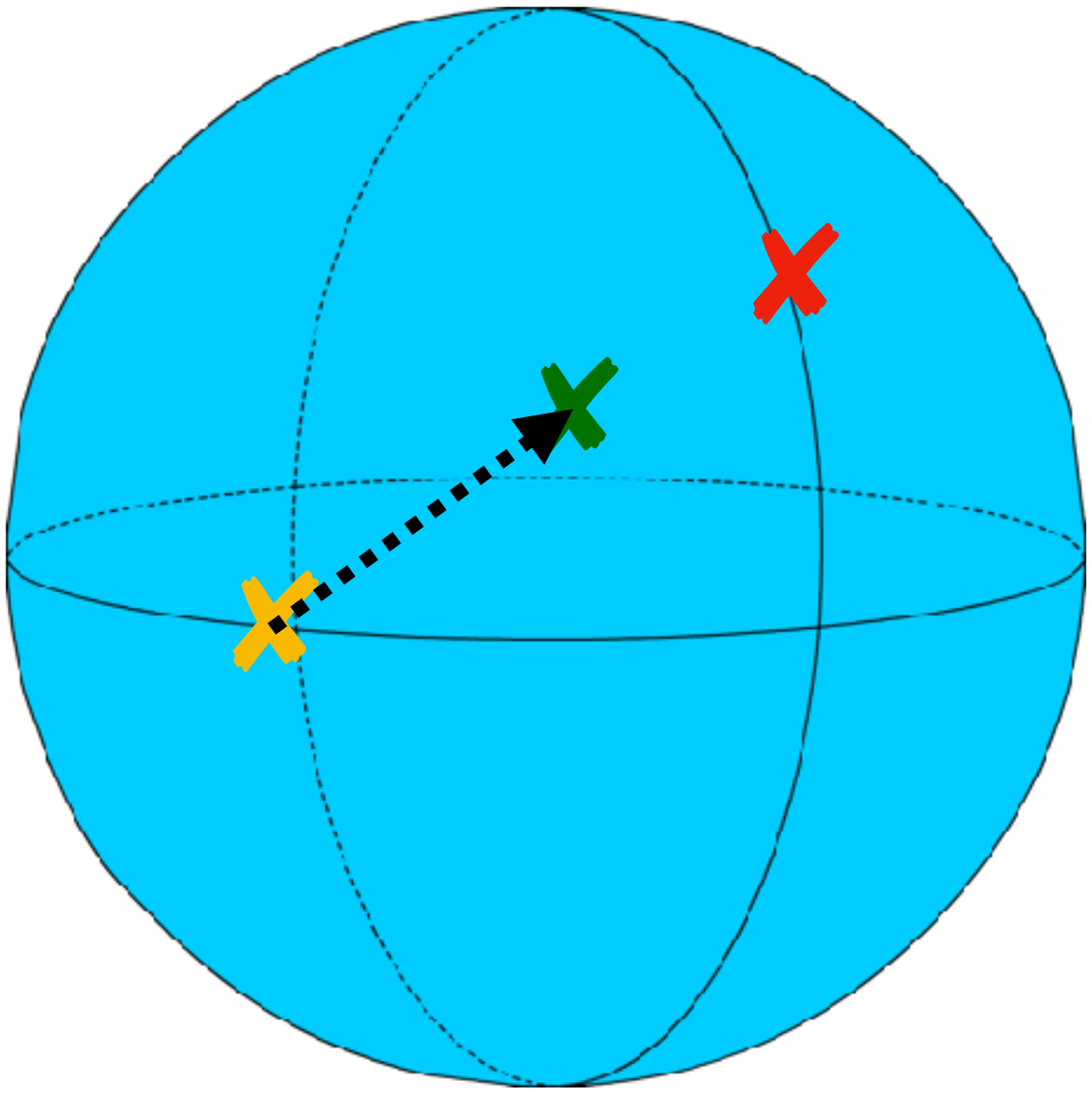

(3, 112, 112)
(96, 53, 53)
(96, 53, 53)
(96, 53, 53)
(96, 18, 18)
(256, 14, 14)
(256, 14, 14)
(256, 7, 7)
(512, 7, 7)
(512, 7, 7)
(512, 7, 7)
(512, 7, 7)
(512, 7, 7)
(512, 7, 7)
(512, 3, 3)
(2048)
(2048)
(2048)
(128)
(128)
Input
conv
relu
lrn
pool
conv
relu
pool
conv
relu
conv
relu
conv
relu
pool
fc
relu
dropout
fc
l2norm

a p n

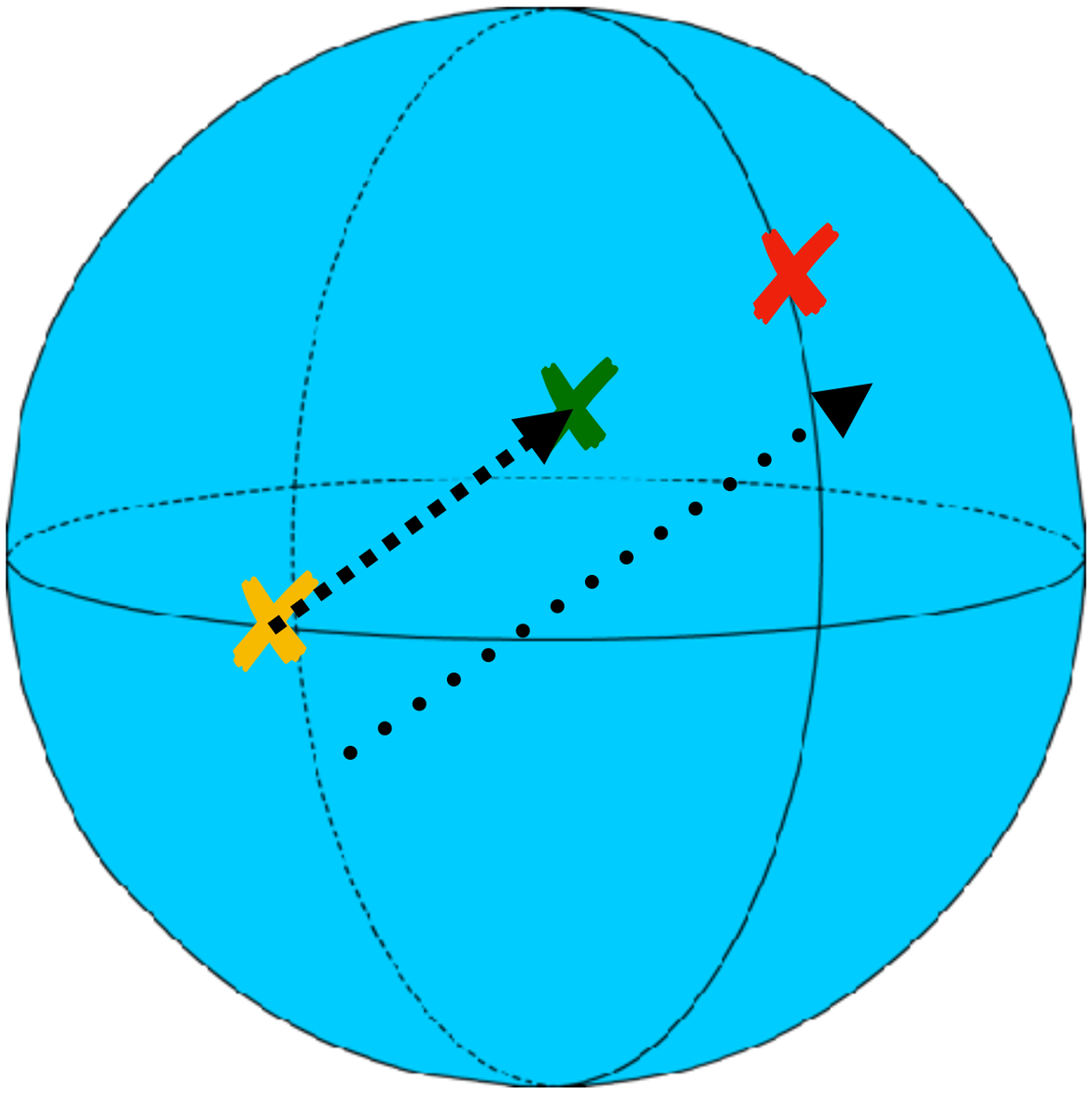




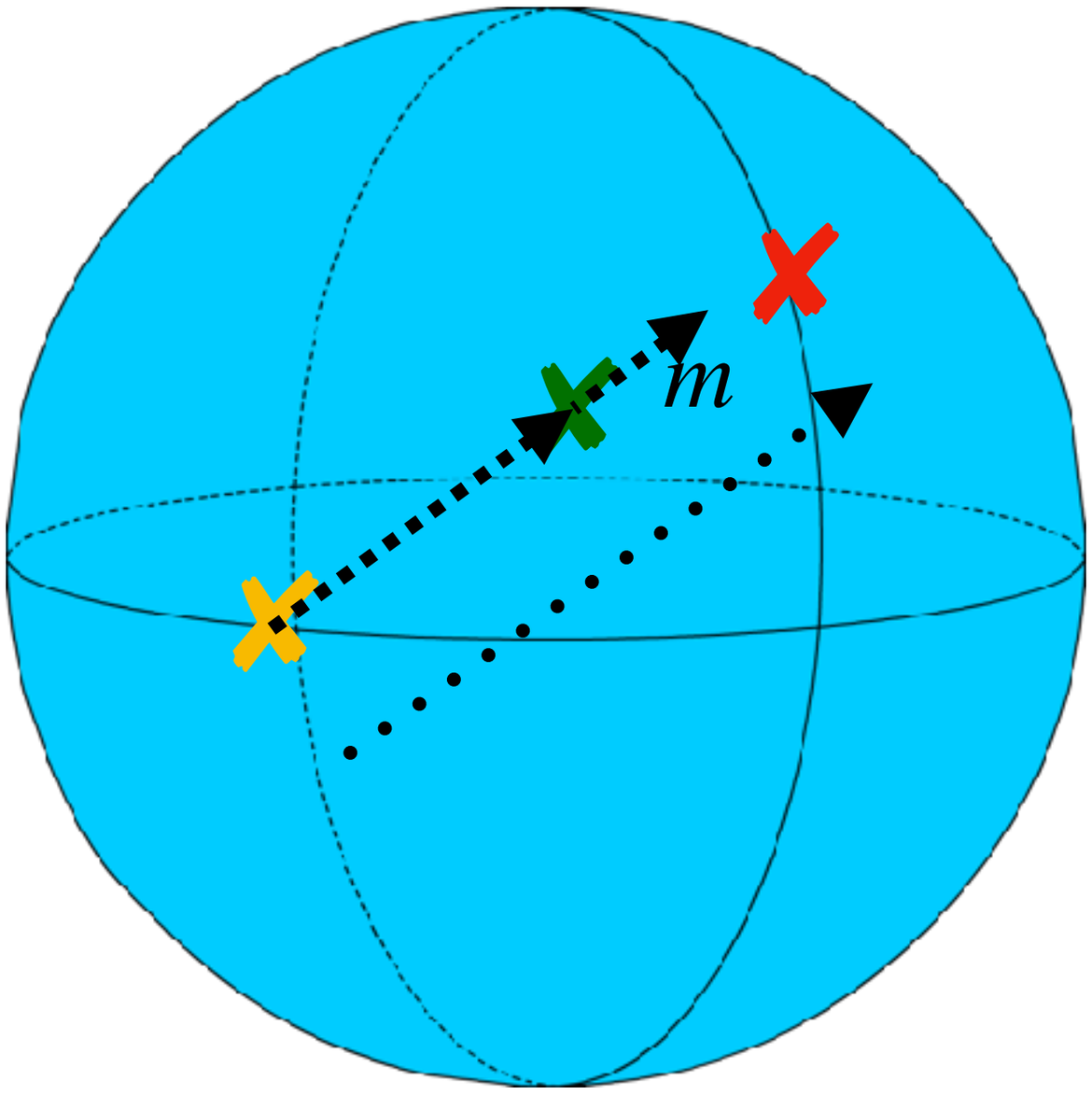
$$\mathcal{L}_{triplet} = \left[||f(a) - f(p)||_2^2 - ||f(a) - f(n)||_2^2 + m \right]_+$$



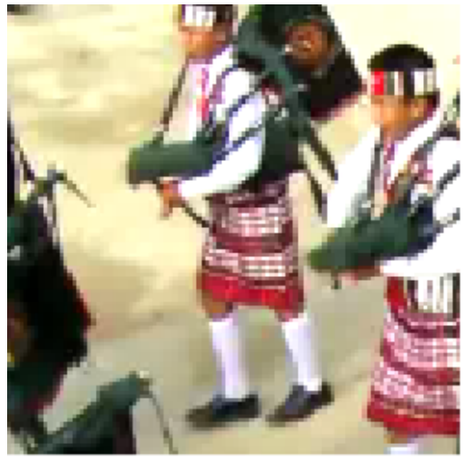
$$\mathcal{L}_{triplet} = \left[||f(a) - f(p)||_2^2 - ||f(a) - f(n)||_2^2 + m\right]_+$$

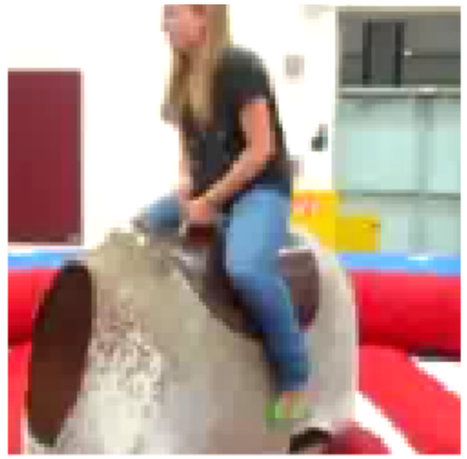


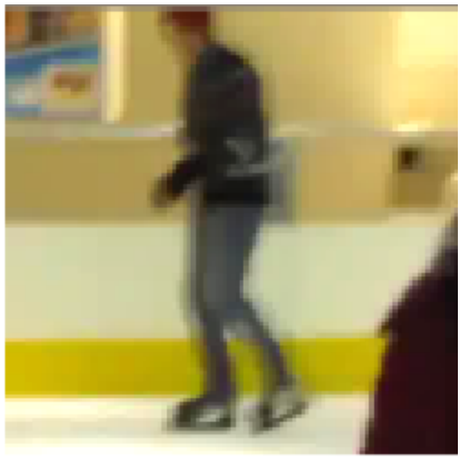
$$\mathcal{L}_{triplet} = \left[||f(a) - f(p)||_2^2 - ||f(a) - f(n)||_2^2 + m \right]_+$$

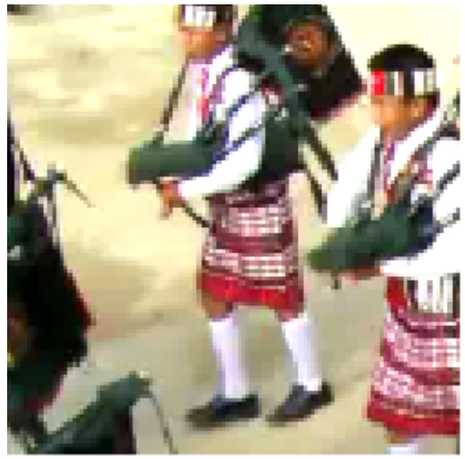
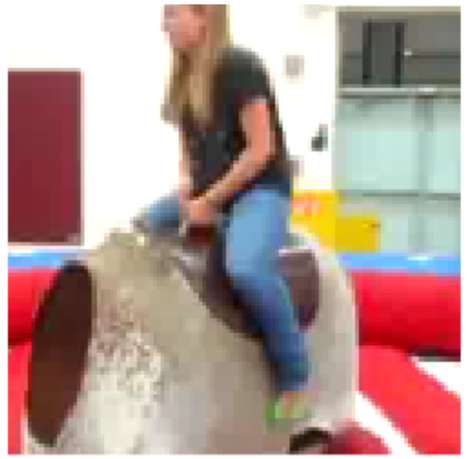


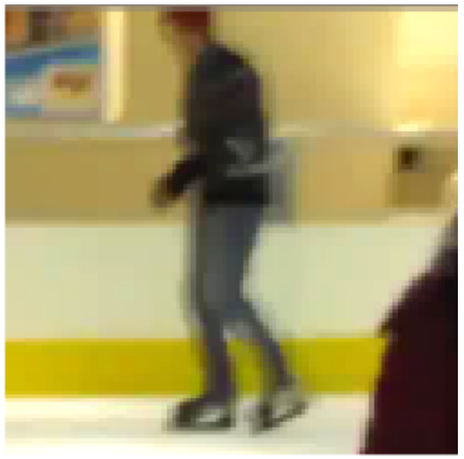
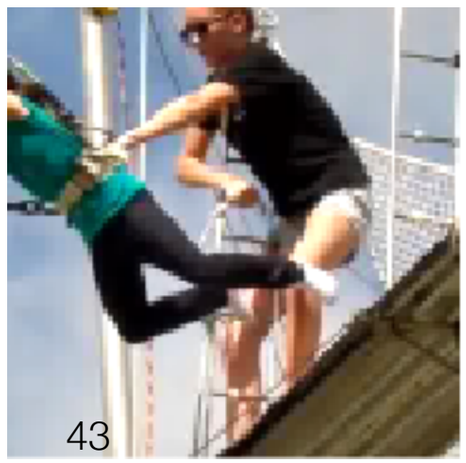
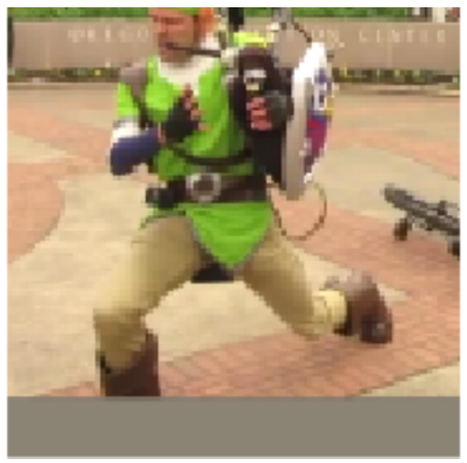
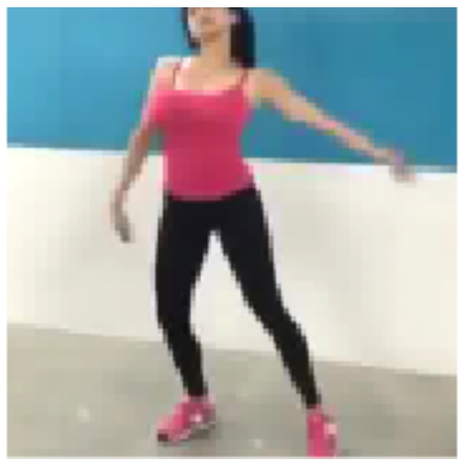
$$\mathcal{L}_{triplet} = \left[||f(a) - f(p)||_2^2 - ||f(a) - f(n)||_2^2 + m\right]_+$$

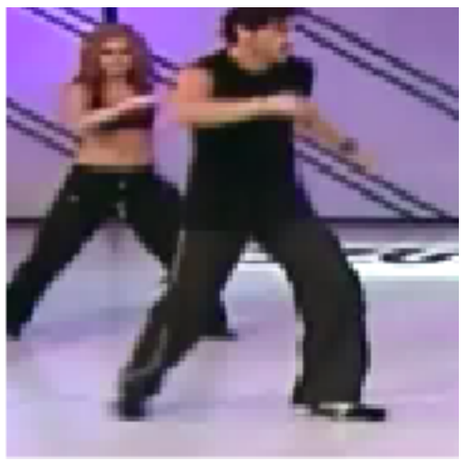


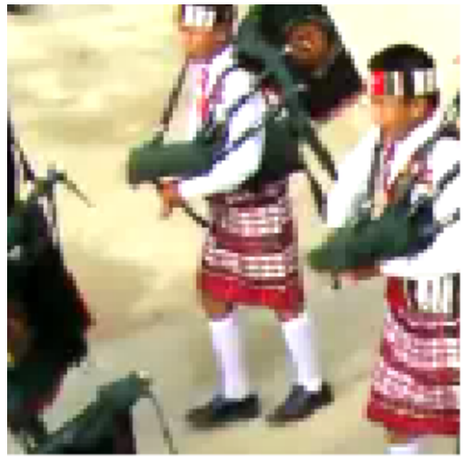
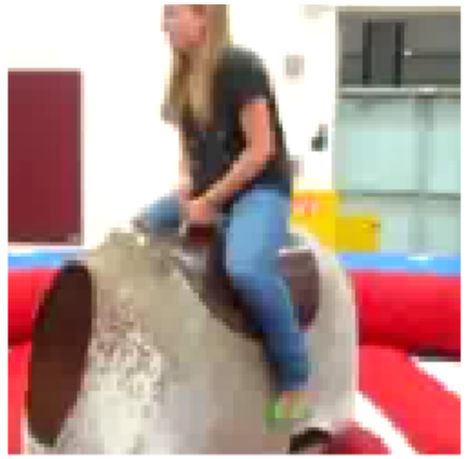


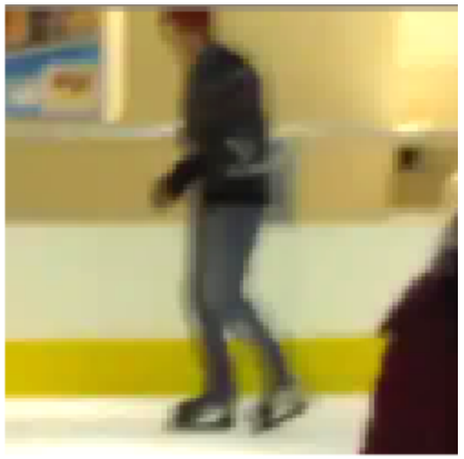


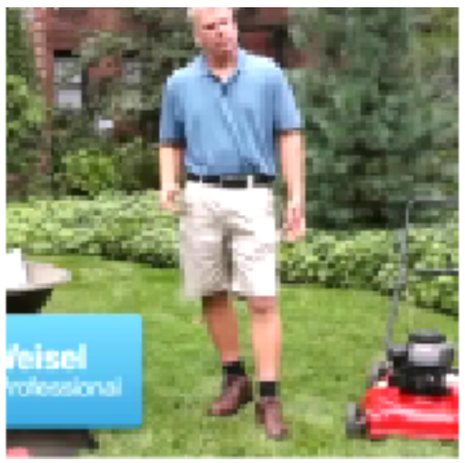
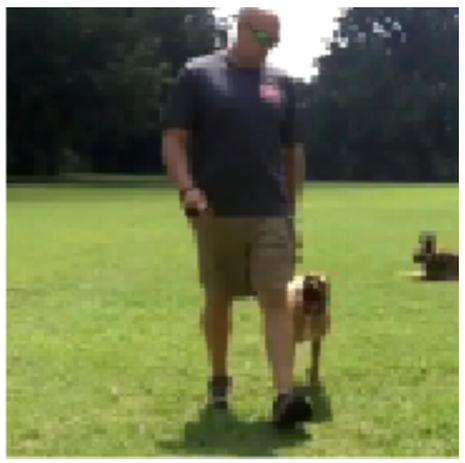






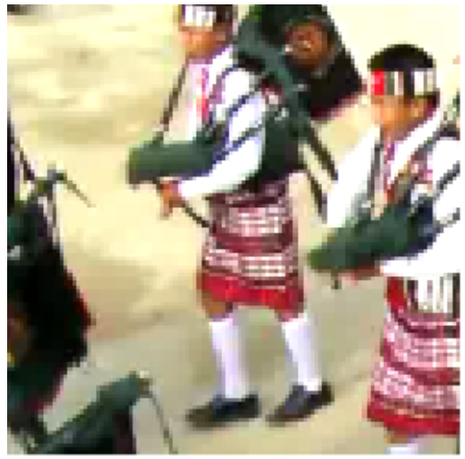
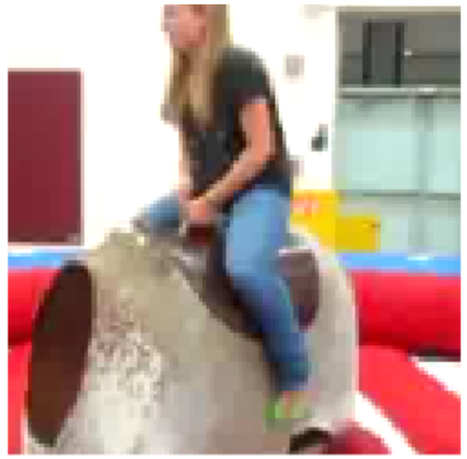


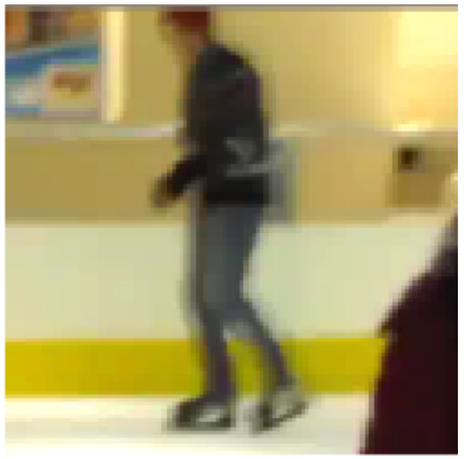
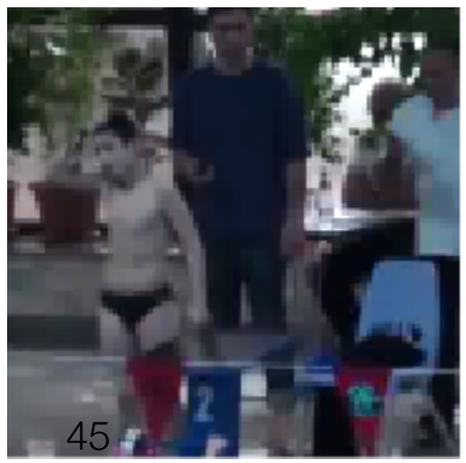



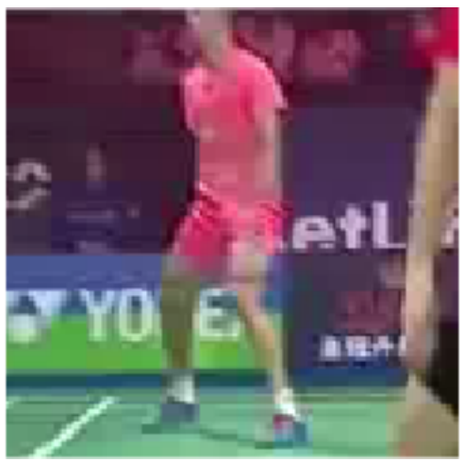






Weisel
Professional


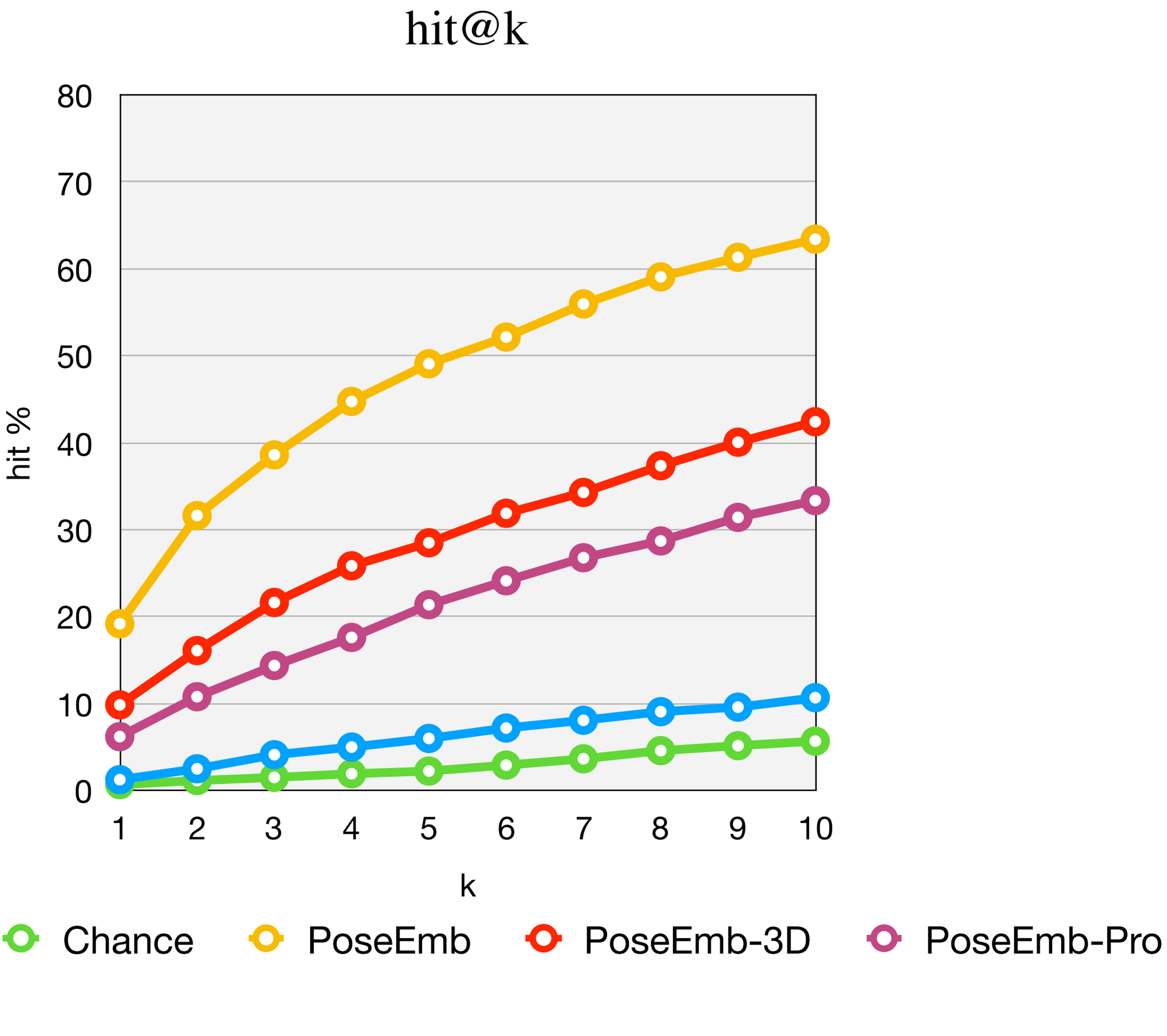



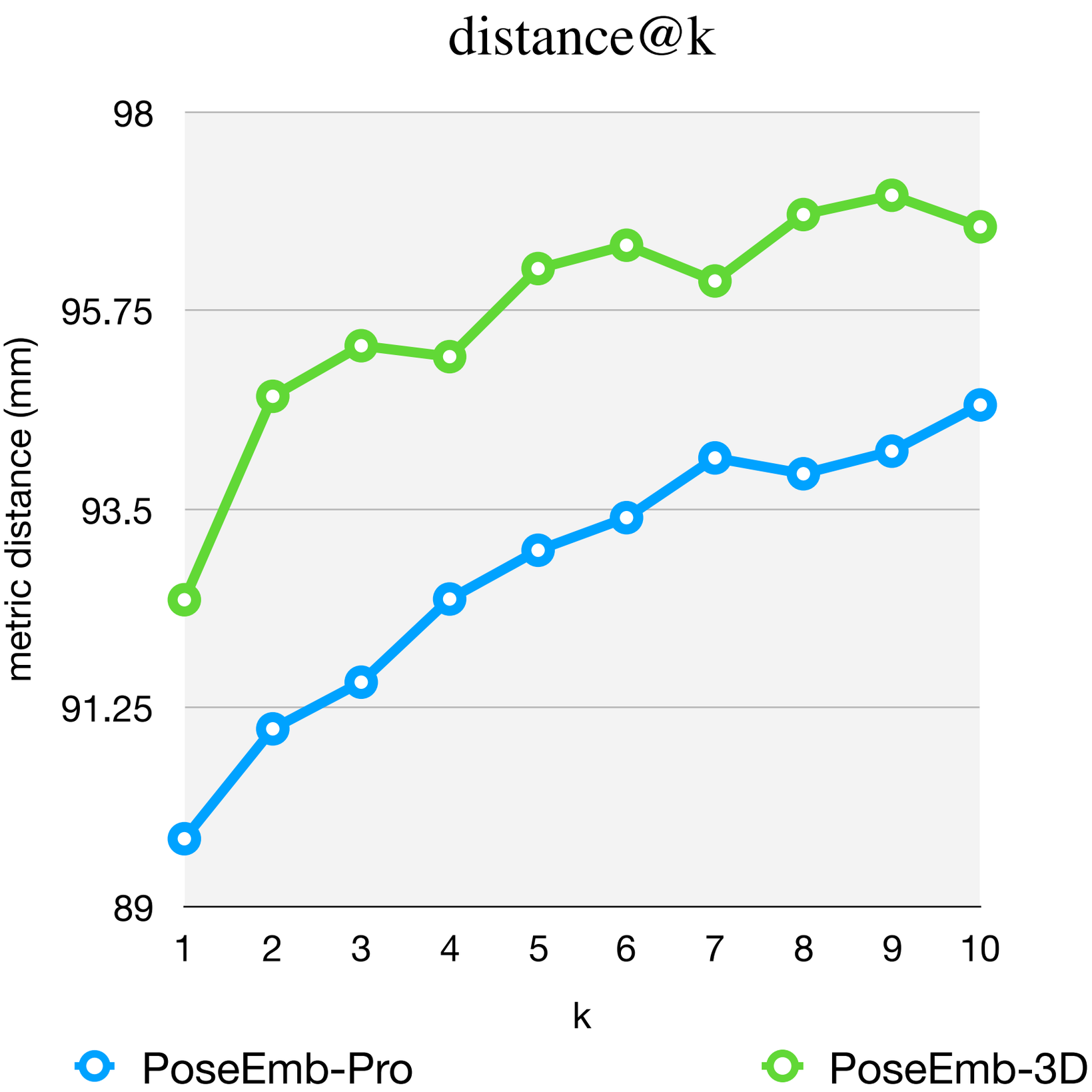

$$hit@k = \frac{1}{N}\sum_{i=1}^{N} hit_k(R^i_{1...k})$$

$$hit_k(X) = \begin{cases} 1, & \text{if } |X \cap T_{50}| > 0 \\ 0, & \text{otherwise} \end{cases}$$

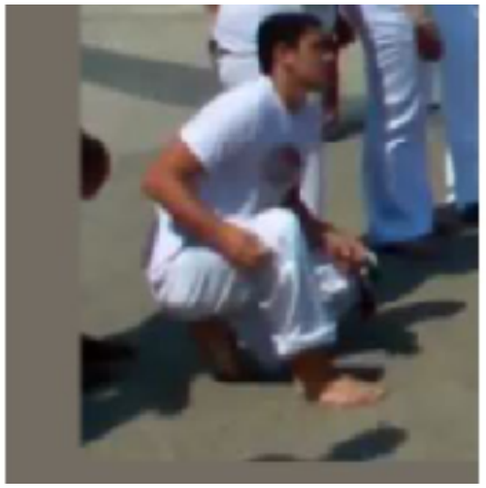

hit@k
80
70
60
50
40
30
20
10
0
hit %
1
2
3
4
5
6
7
8
9
10
k
VGG-S
Chance
PoseEmb
PoseEmb-3D
PoseEmb-Pro

$$distance@k = \frac{1}{N} \sum_{i=1}^{N} distance_k(R^i_{1...k})$$

$$distance_k(X) = \frac{1}{N} \sum_{i=1}^{N} \delta(X_i, q)$$

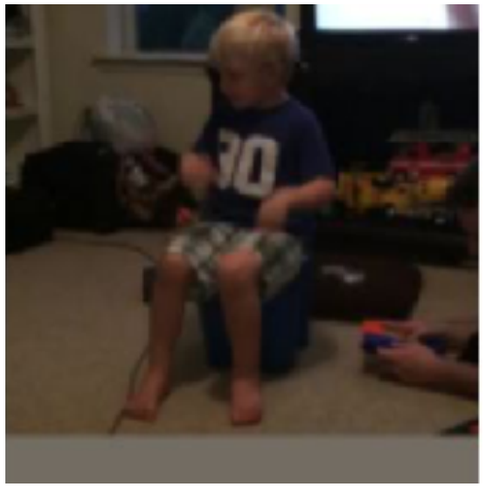

distance@k
98
95.75
93.5
91.25
89
metric distance (mm)
1
2
3
4
5
6
7
8
9
10
k
PoseEmb-Pro
PoseEmb-3D

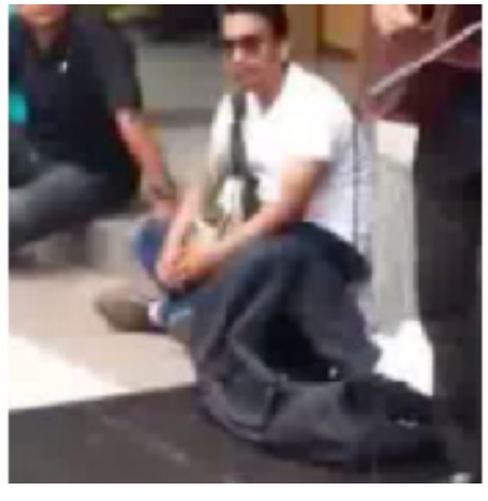
2D

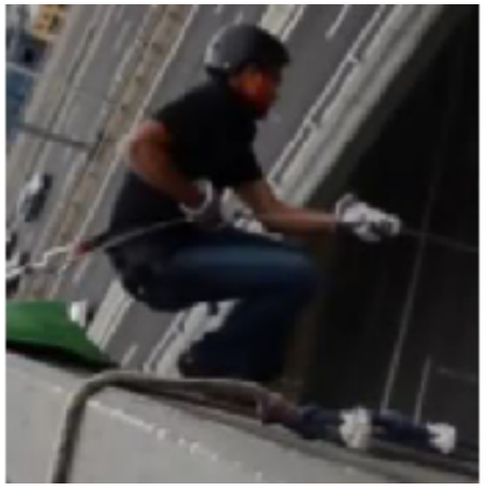


3D

Procrustes







Language Querying

“Put your hands above your head!”

Posebits for Monocular Human Pose Estimation

Gerard Pons-Moll
MPI for Intelligent Systems
Tübingen, Germany
gerard.pons.moll@tue.mpg.de

David J. Fleet
University of Toronto
Toronto, Canada
fleet@cs.toronto.edu

Bodo Rosenhahn
Leibniz University of Hannover
Hannover, Germany
rosenhahn@tnt.uni-hannover.de

Abstract

We advocate the inference of qualitative information about 3D human pose, called posebits, *from images. Posebits represent boolean geometric relationships between body parts (*e.g. *left-leg in from of right-leg or hands close to each other). The advantages of posebits as a mid-level representation are 1) for many tasks of interest, such qualitative pose information may be sufficient (*e.g. *semantic image retrieval), 2) it is relatively easy to annotate large image corpora with posebits, as it simply requires answers to yes/no questions; and 3) they help resolve challenging pose ambiguities and therefore facilitate the difficult task of image-based 3D pose estimation. We introduce posebits, a posebit database, a method for selecting useful posebits for pose estimation and a structural SVM model for posebit inference. Experiments show the use of posebits for semantic image retrieval and for improving 3D pose estimation.*

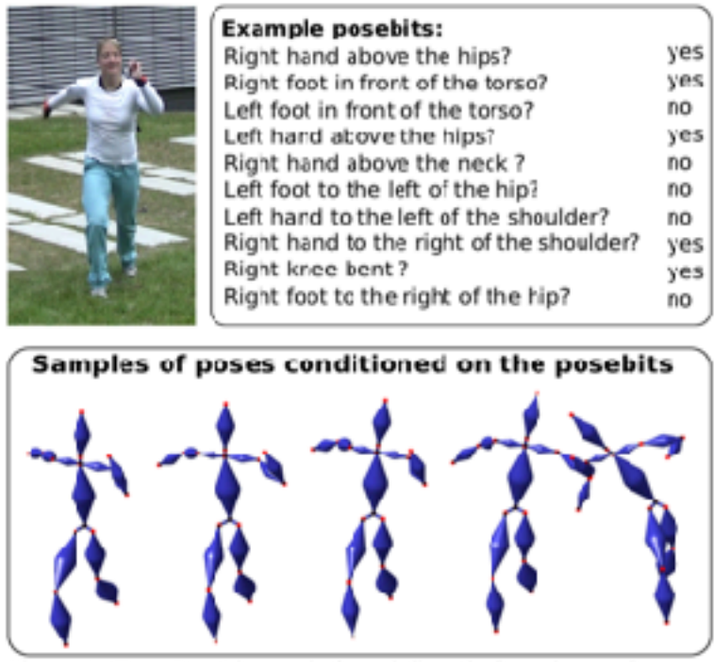


Figure 1. Top: Posebits are inferred directly from image features using a trained classifier. Since posebits consist of simple yes/no questions, images can be easily annotated by humans. They may be useful for many tasks. For example, on the bottom image, we show samples of poses conditioned on the posebits depicted on the top image. By conditioning the poses on posebits uncertainty about the pose is reduced. Notice how the poses are qualitatively very similar to the observed image. In this example we show the ground truth posebyte. Our model also takes into account uncertainty in the estimation of posebits by marginalizing over them.

1. Introduction

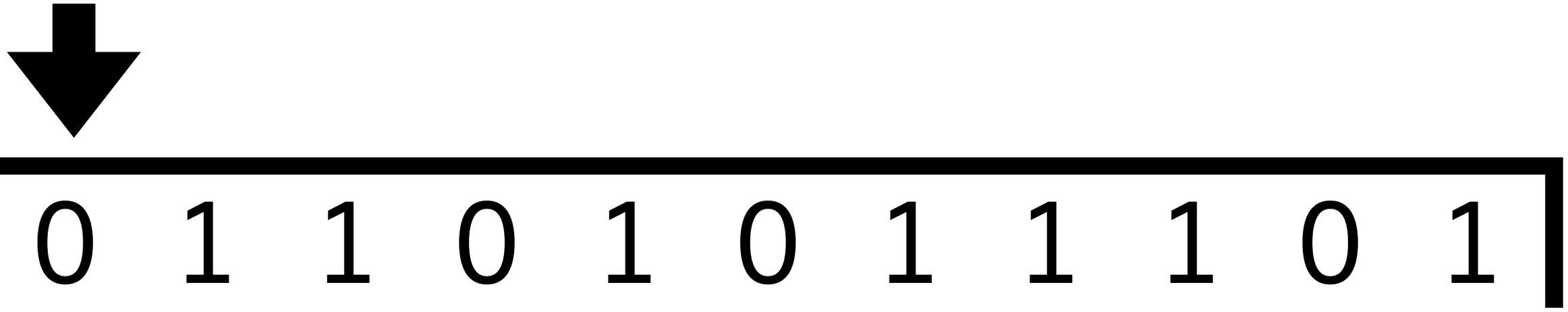
While tremendous effort has focused on the extraction of quantitative 3D human pose from images and video, in this paper we consider the estimation of qualitative pose information, called *posebits*. Posebits are attributes of pose that specify the relative positions or orientations of body parts. They have the advantage that one can easily collect posebit image annotations for training purposes, and they can be reliably inferred from images. Further, they are useful for resolving 3D pose ambiguities and for myriad other tasks where quantitative pose is not required.

The effective use of both generative and discriminative approaches to pose estimation require training data com-

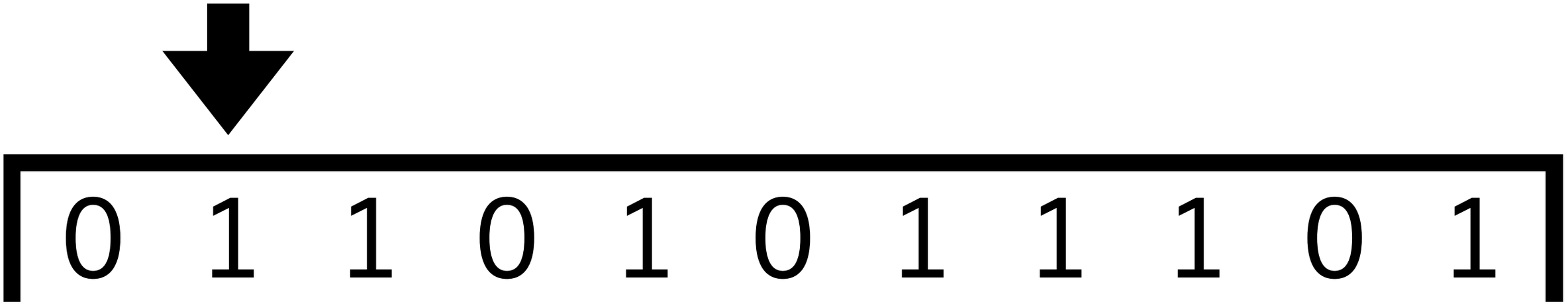
to errors [5, 6].

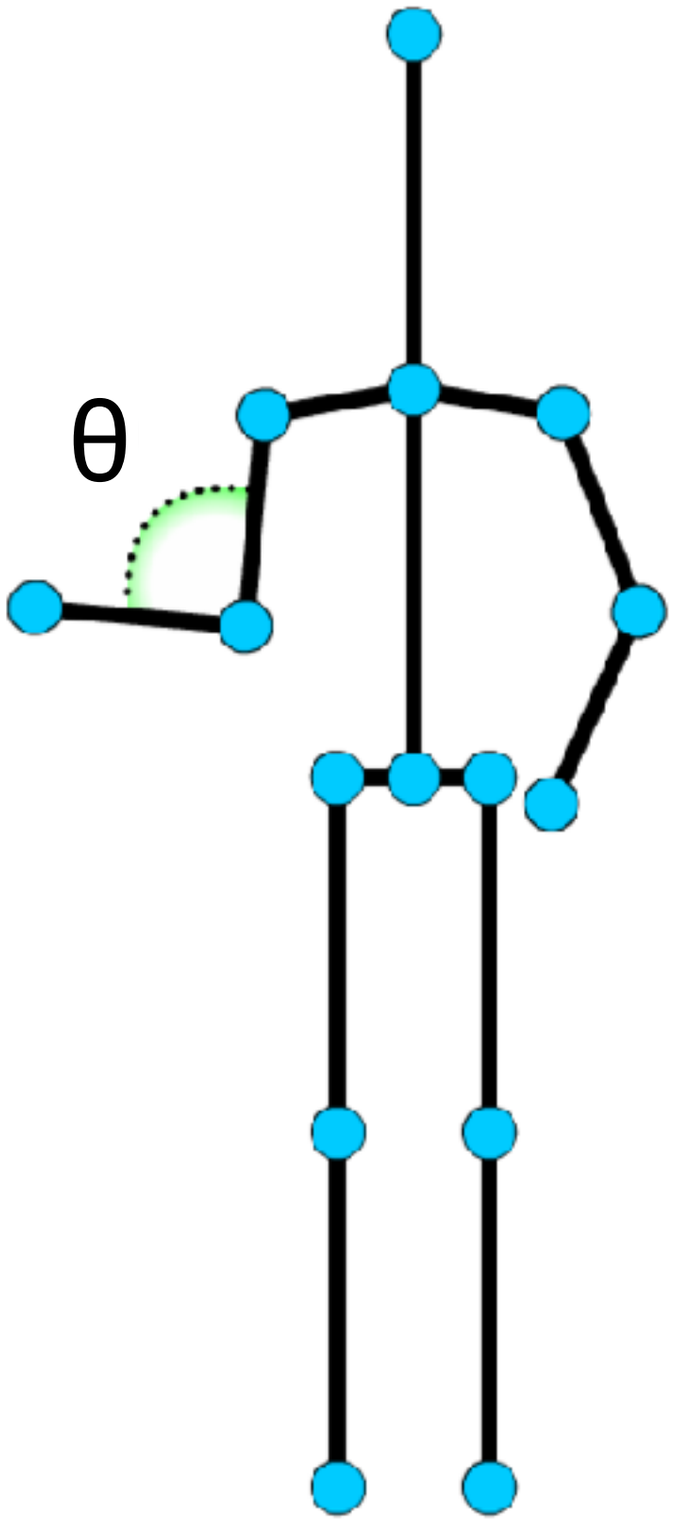
By contrast, it is relatively simple to obtain training data for posebits from human annotations. Indeed, people often

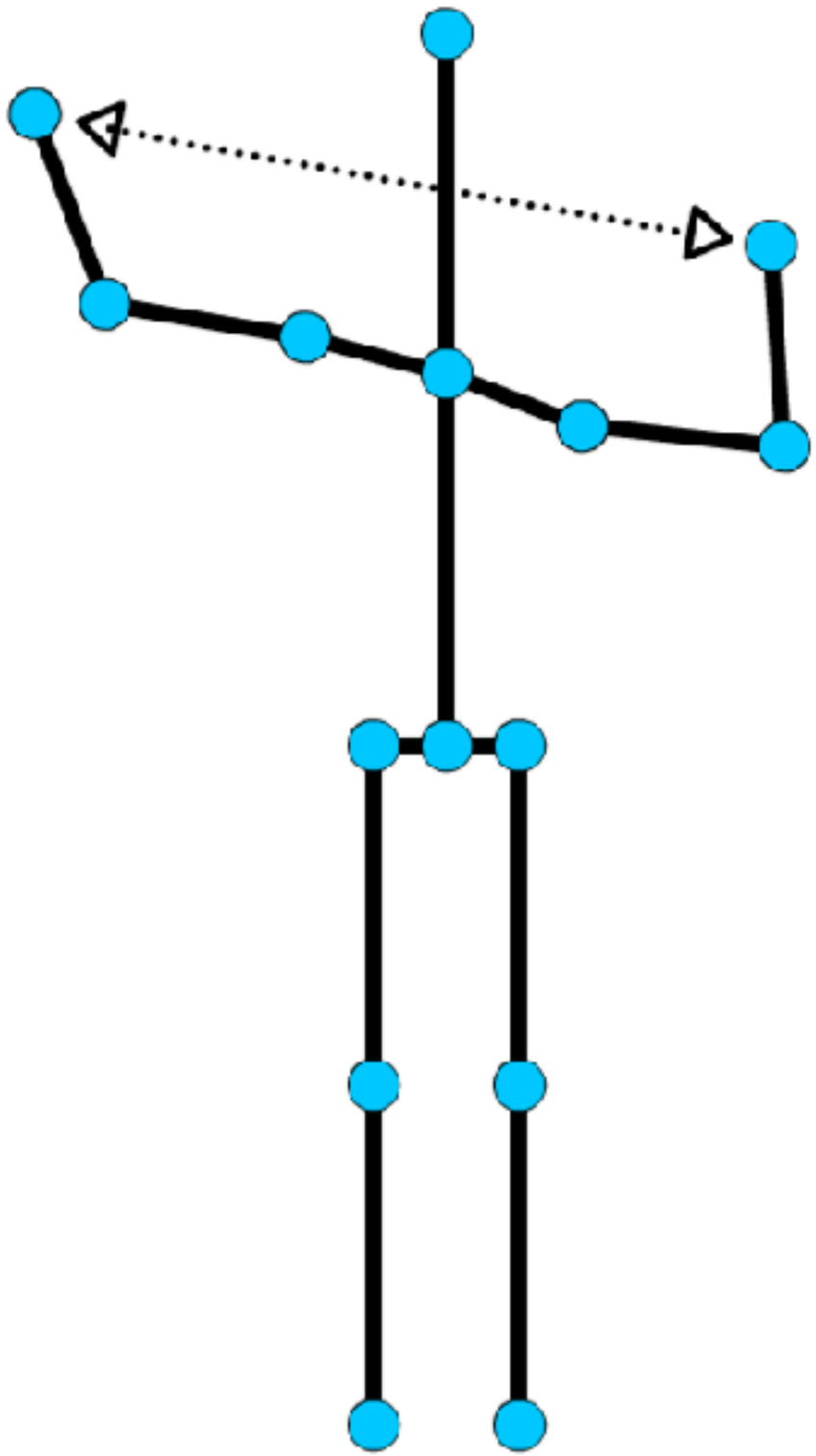
“Right knee **is not** bent”

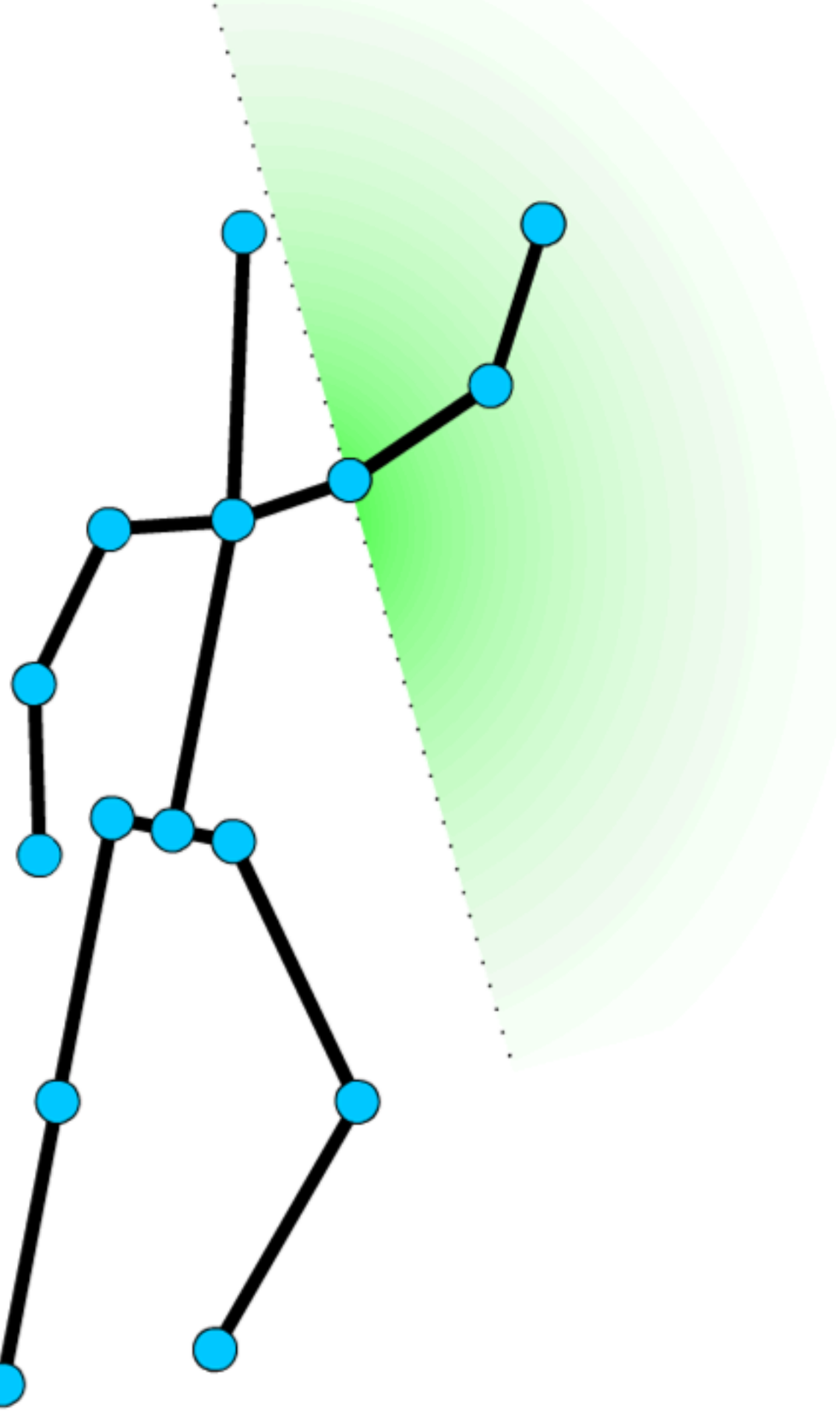


“Left elbow **is** bent”



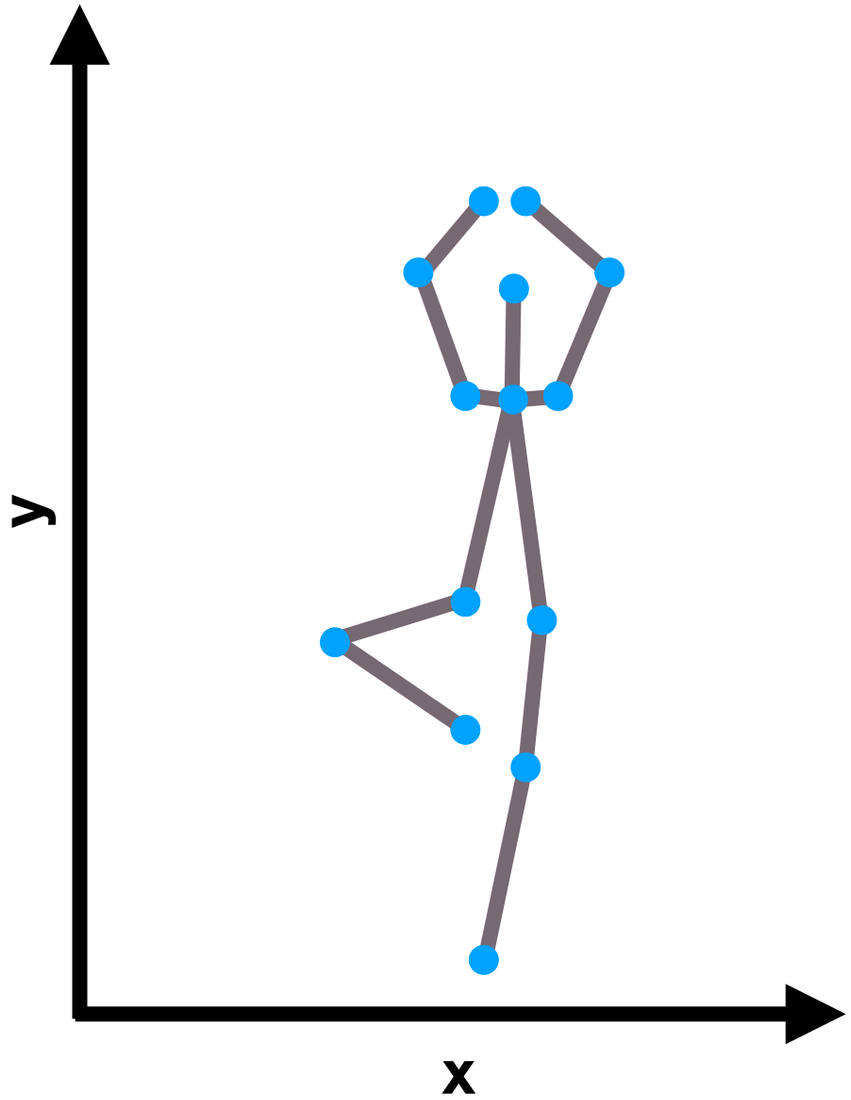

θ

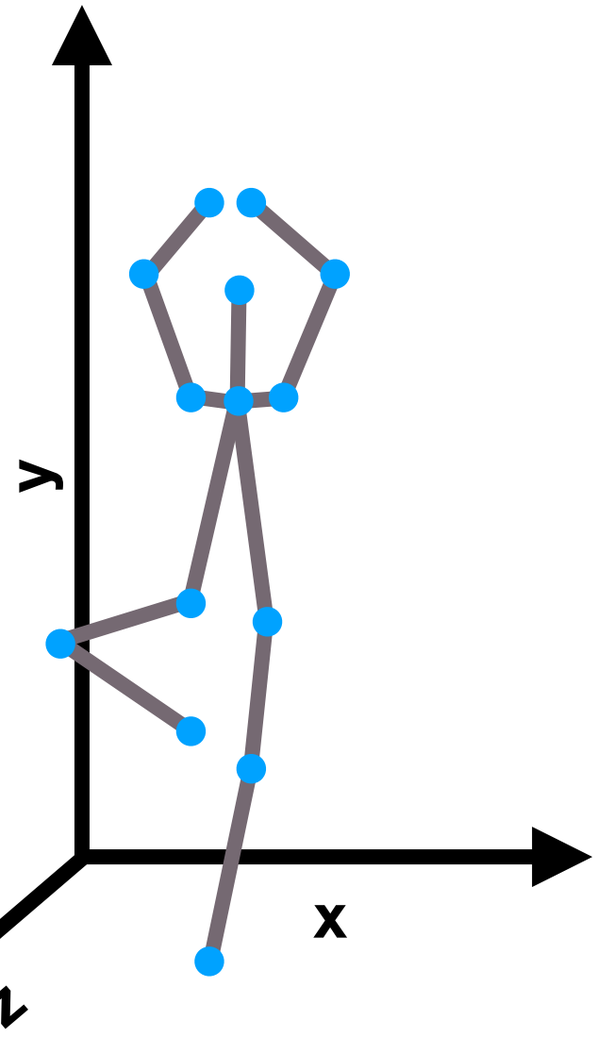




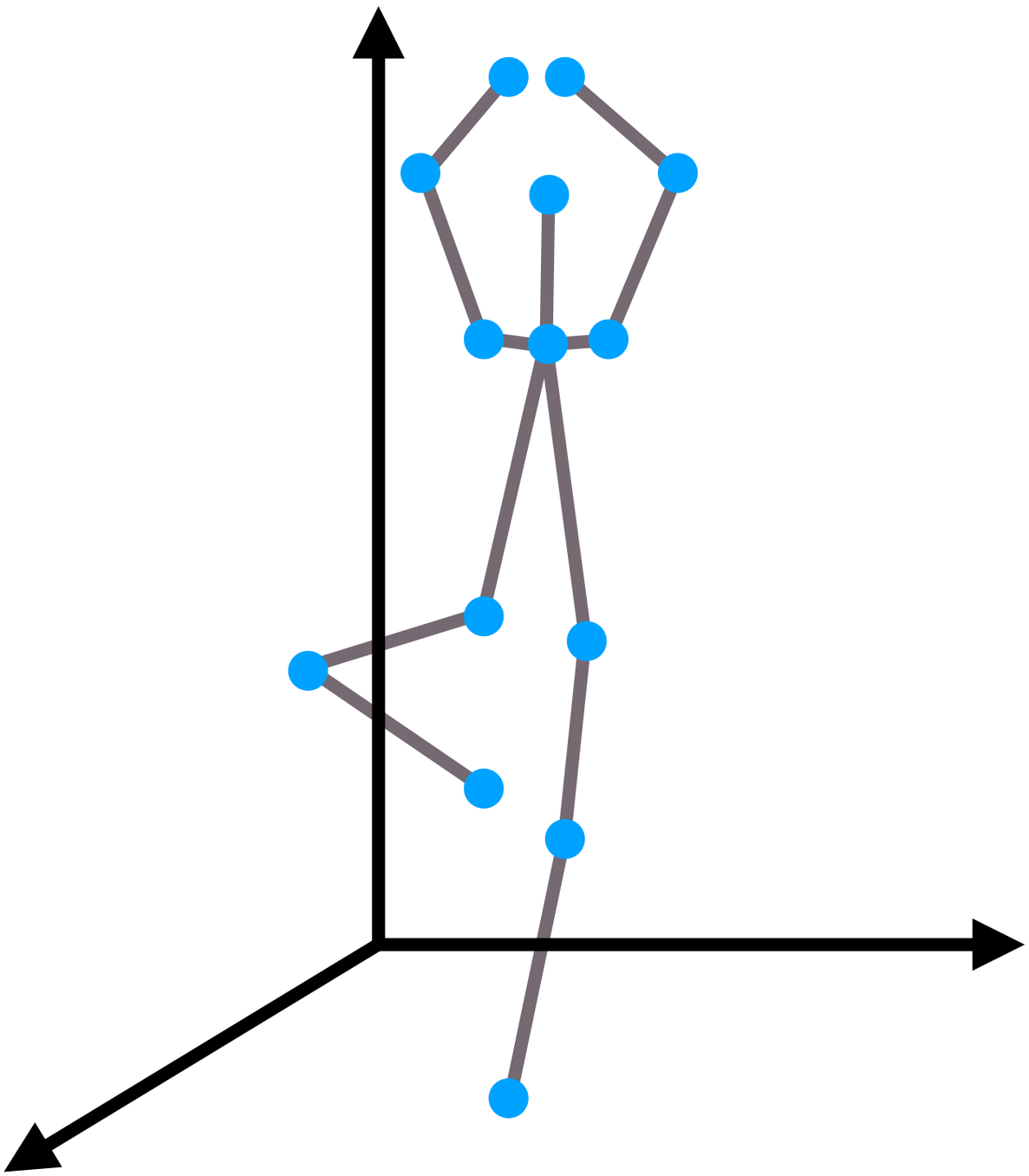
“Put your hands above your head!”

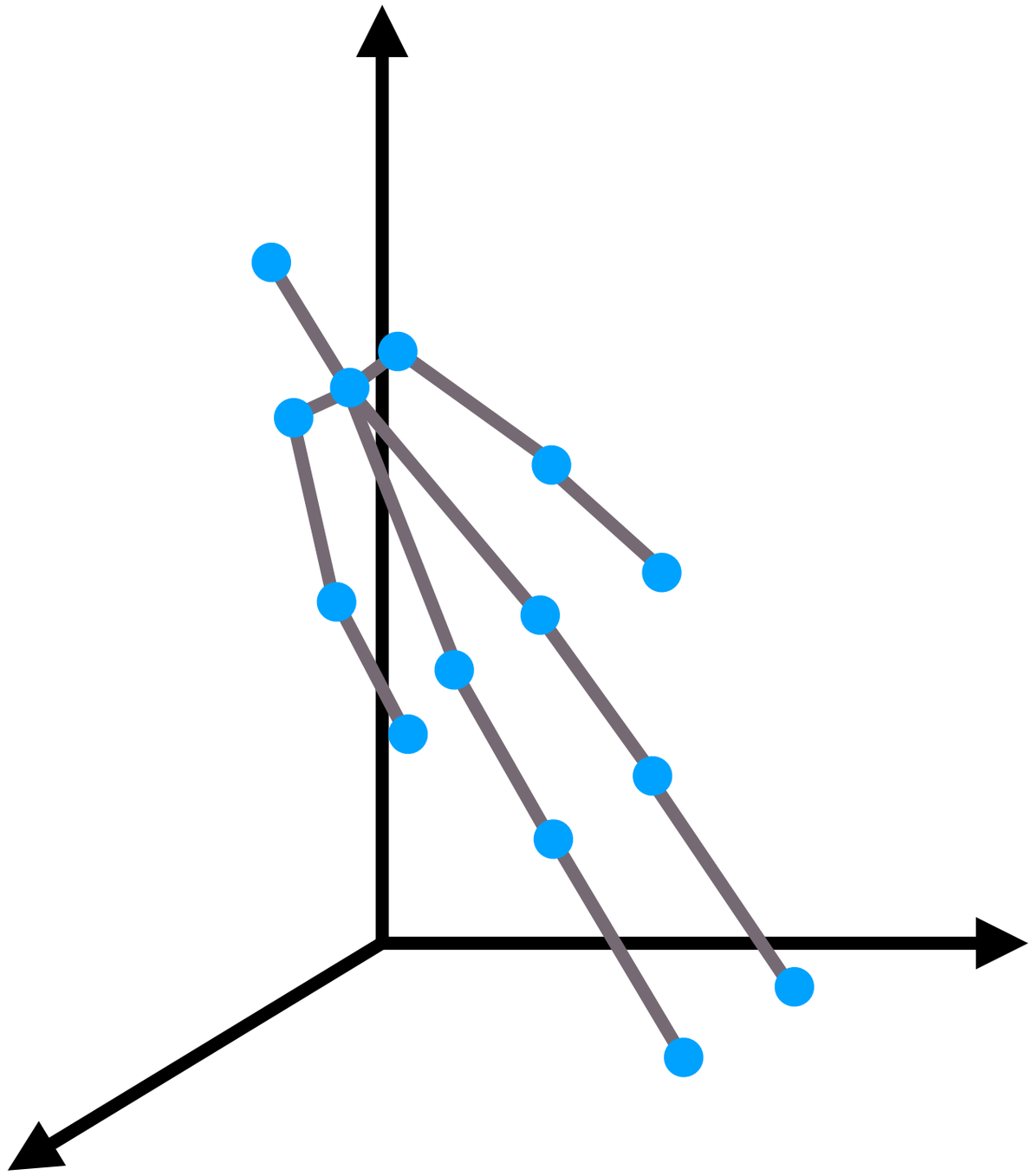
“Put your left hand above your head;
Put your right hand above your head!”





[1 0 0 1 0 1 1 1 0 1 0 1 1]





$$\delta_{hamming}(J_a, J_b) = \frac{1}{P} \sum J_a \oplus J_b$$

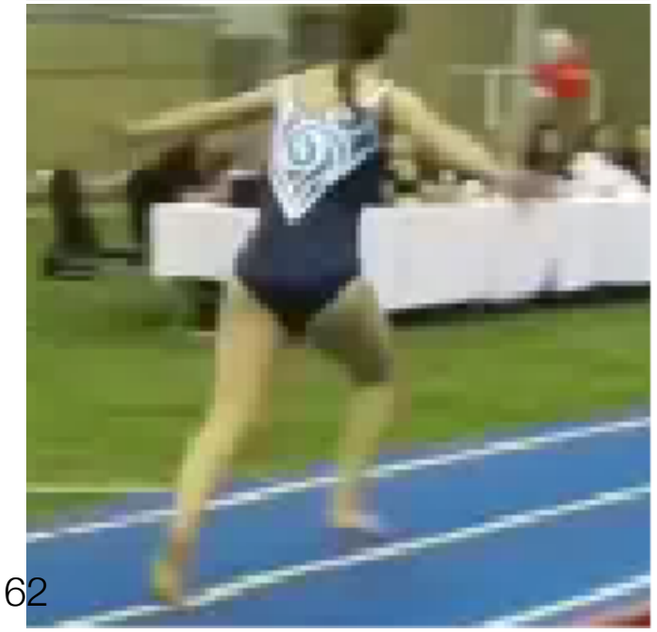
$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \qquad \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

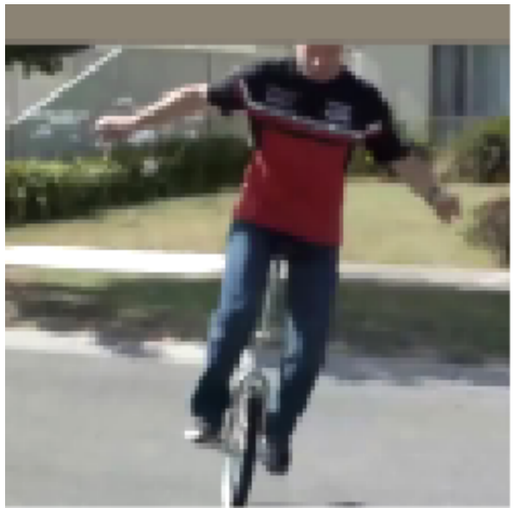
$$\delta_{hamming}(J_a, J_b) = \frac{1}{P} \sum J_a \oplus J_b$$

$$
\begin{bmatrix}
1 \oplus 0 \\
1 \oplus 1 \\
1 \oplus 0 \\
\vdots \\
1 \oplus 1 \\
0 \oplus 1 \\
1 \oplus 0 \\
1 \oplus 1
\end{bmatrix}
$$

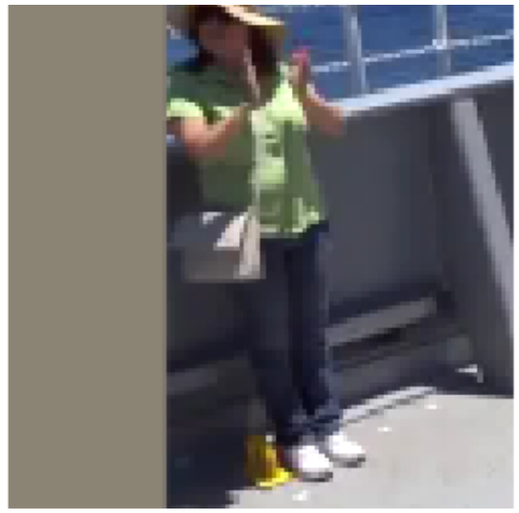
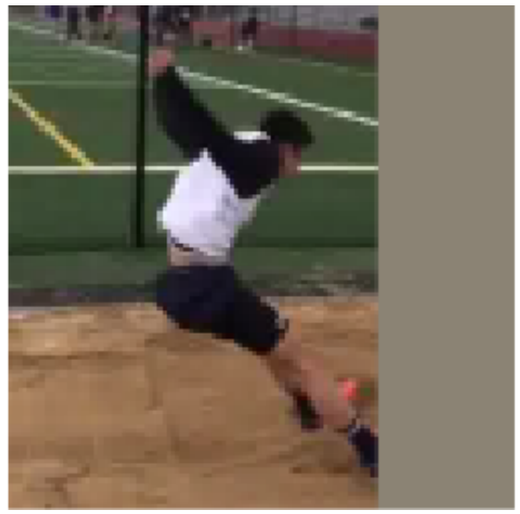
$$\delta_{hamming}(J_a, J_b) = \frac{1}{P} \sum J_a \oplus J_b$$

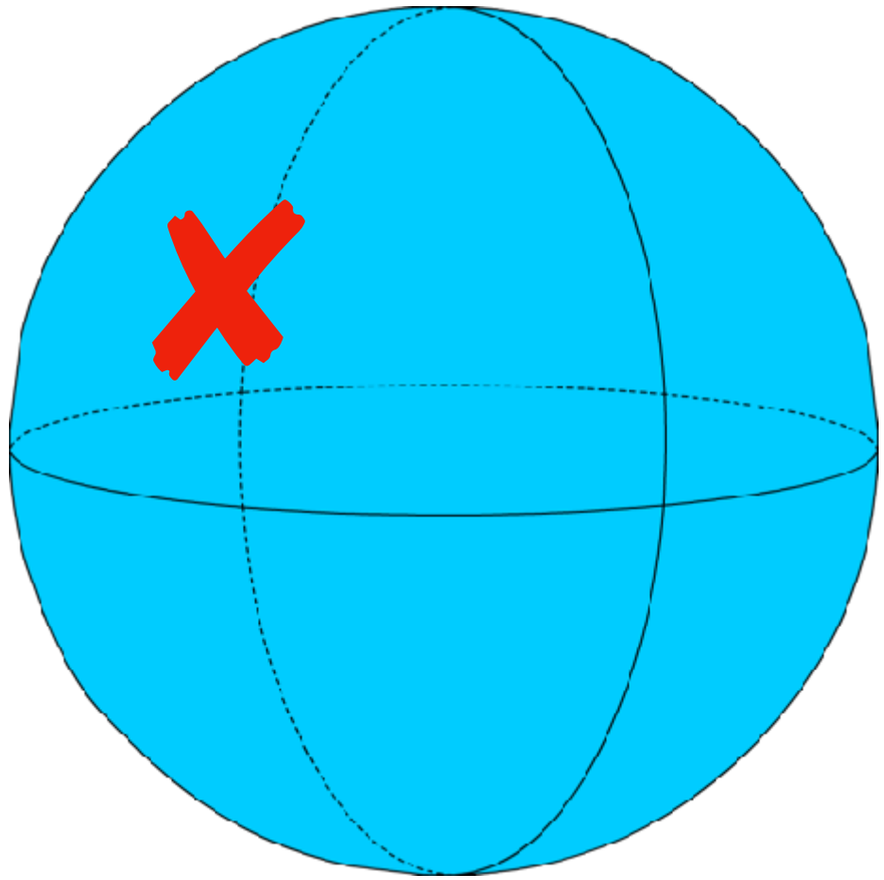


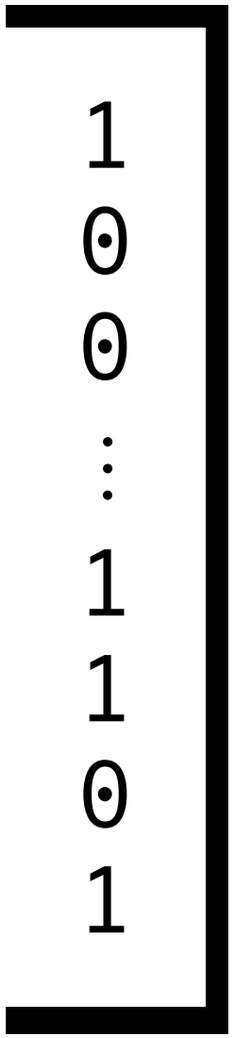


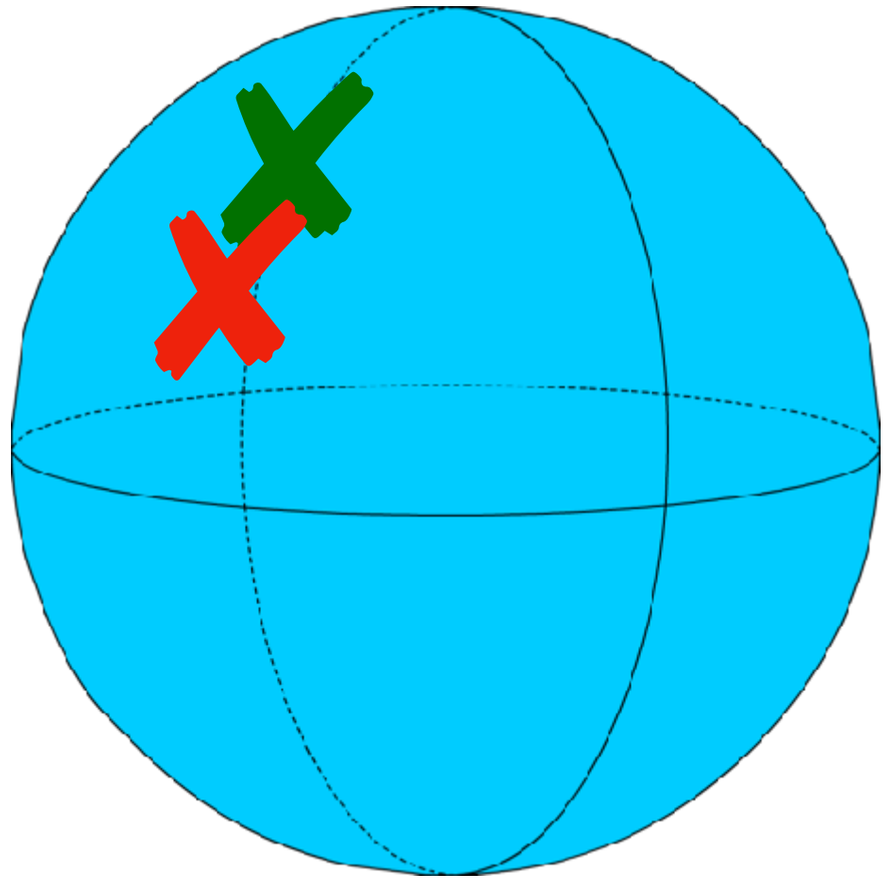



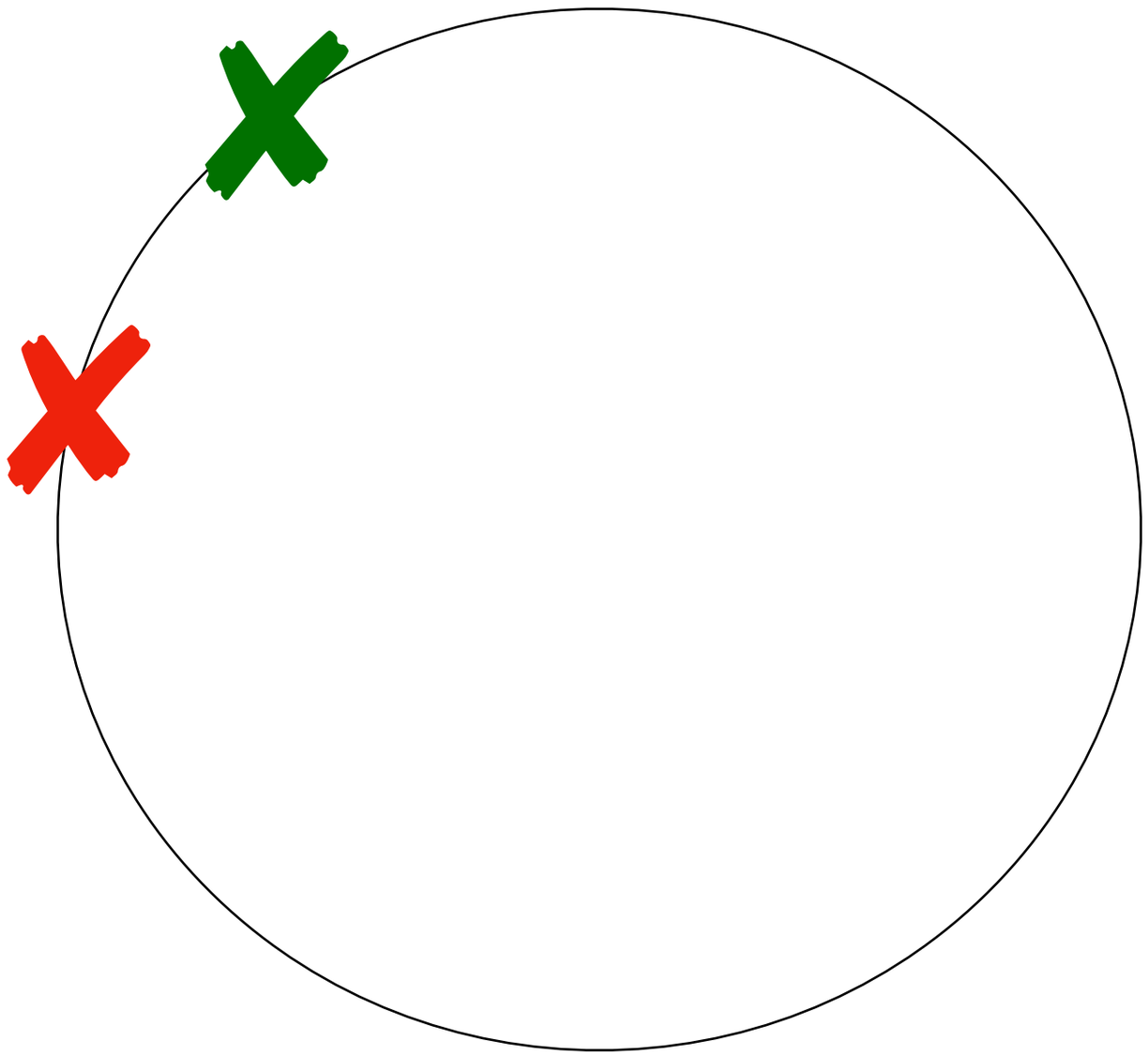


clasmith



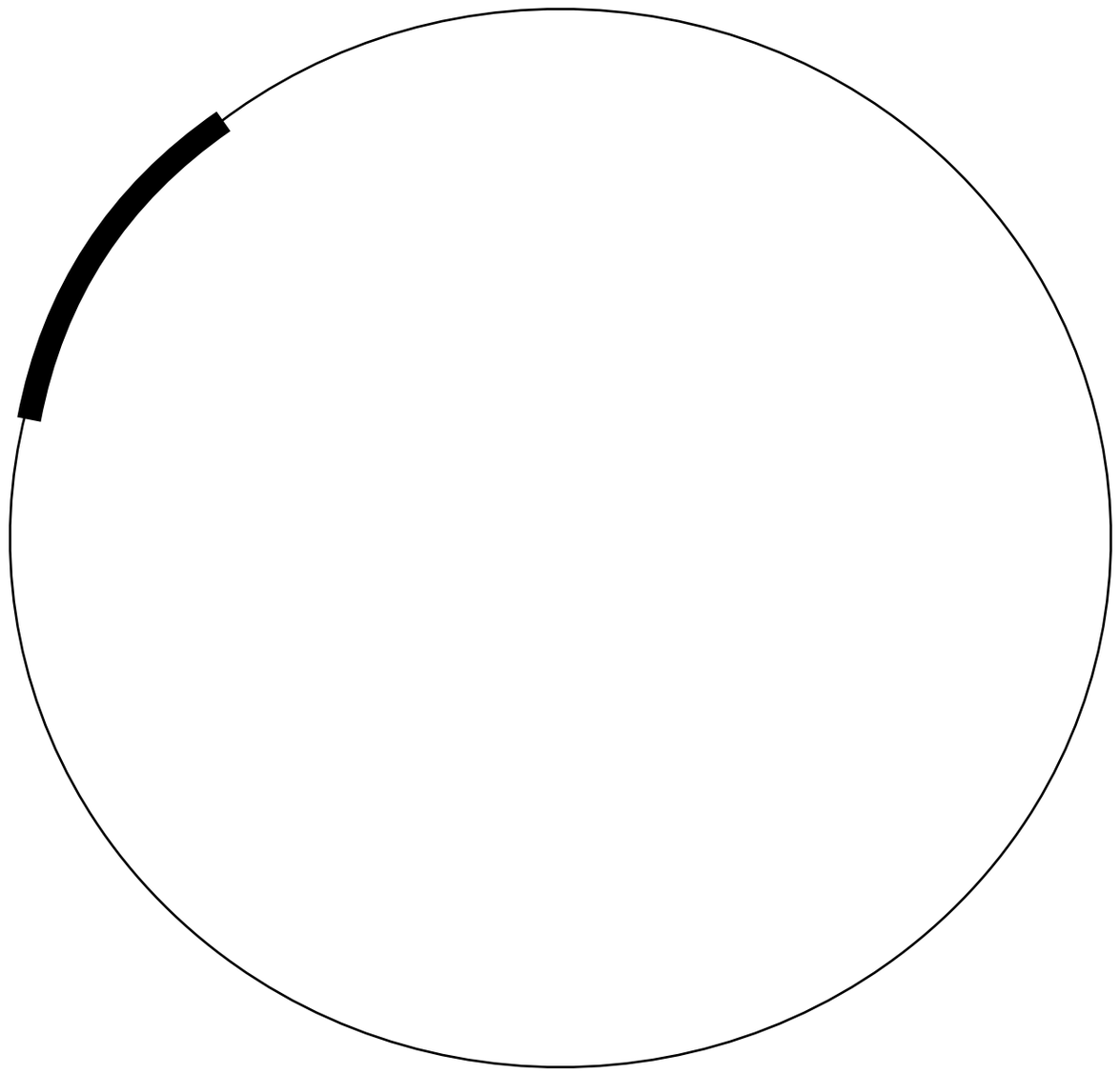




1
0
0
…
1
1
0
1

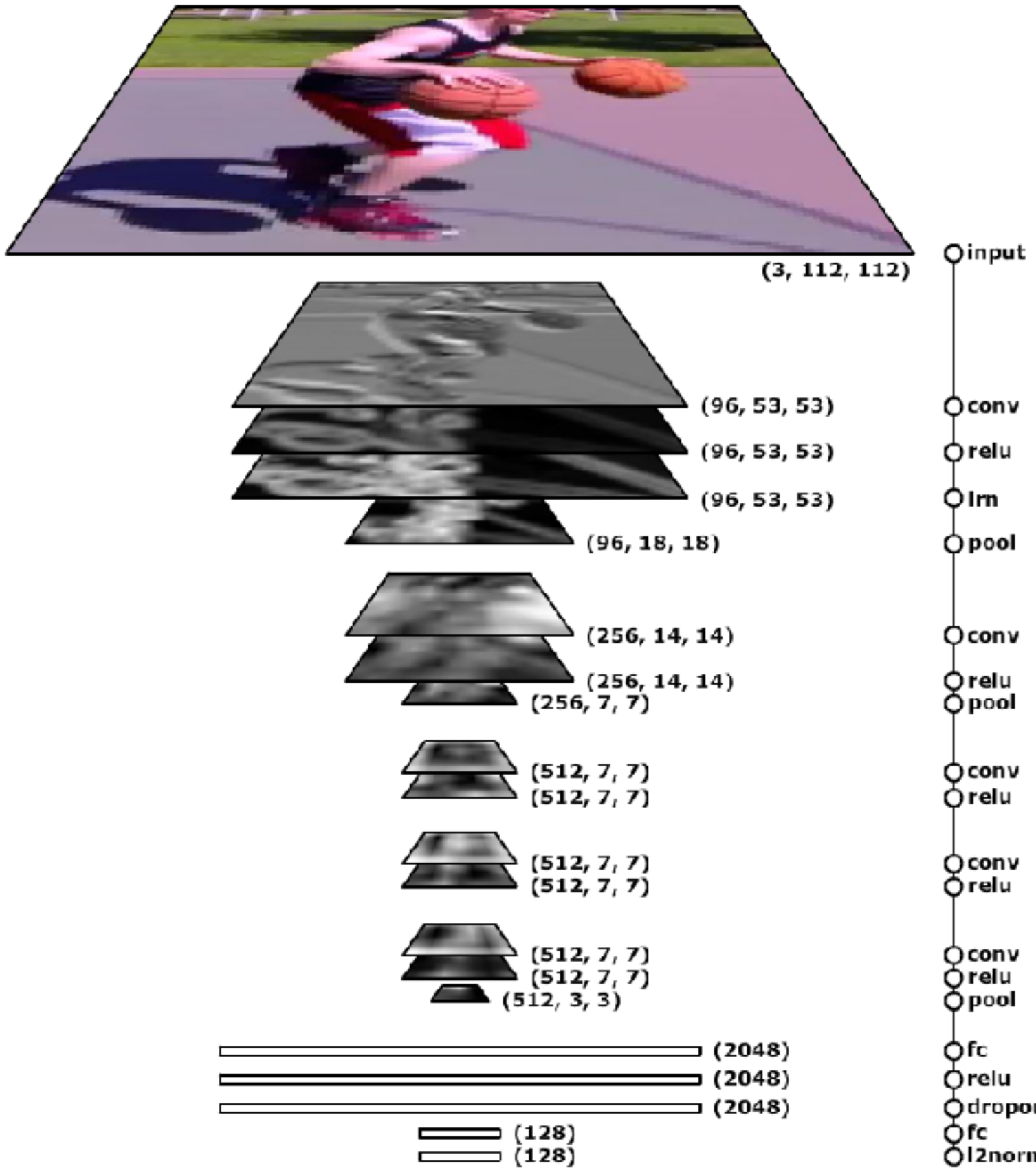




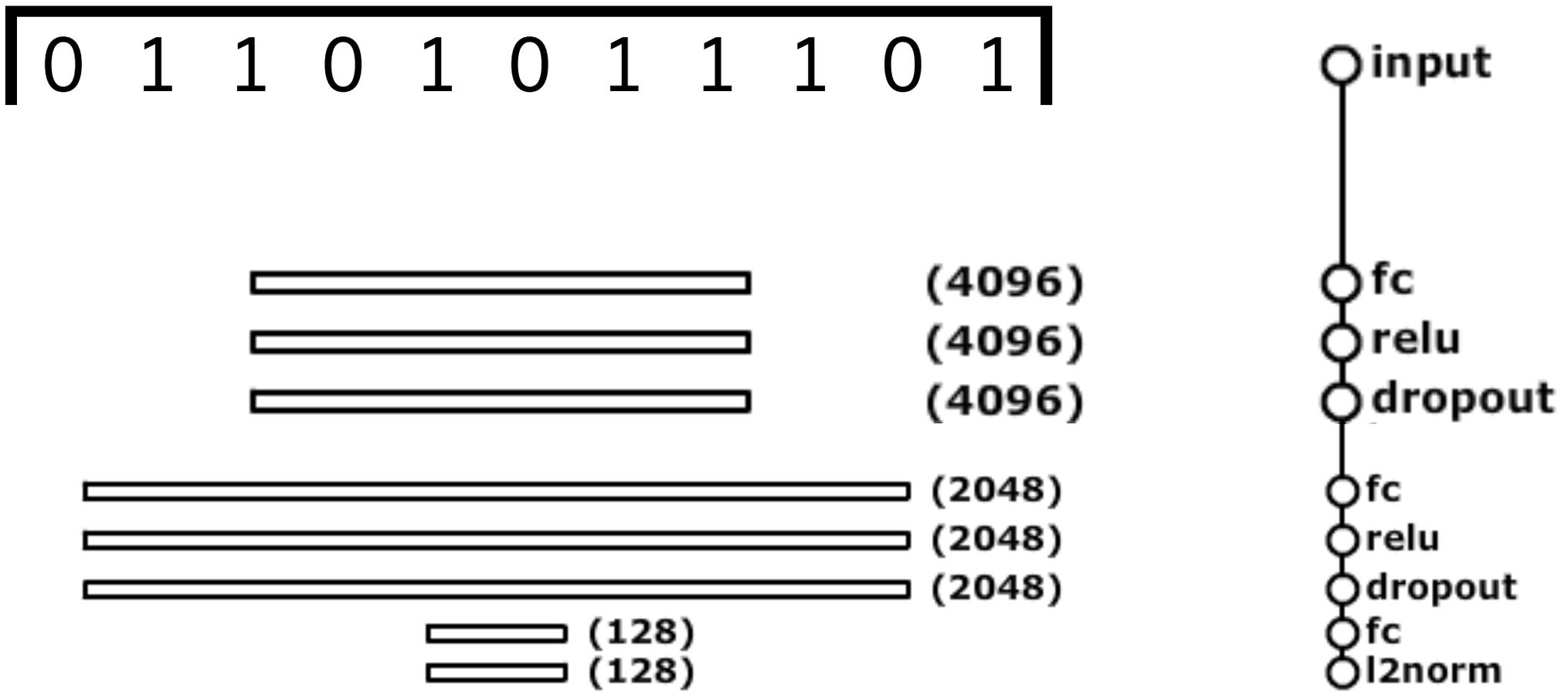
$$\mathcal{L}_c = 1 - f(e_{image}) \cdot g(e_{pose})$$

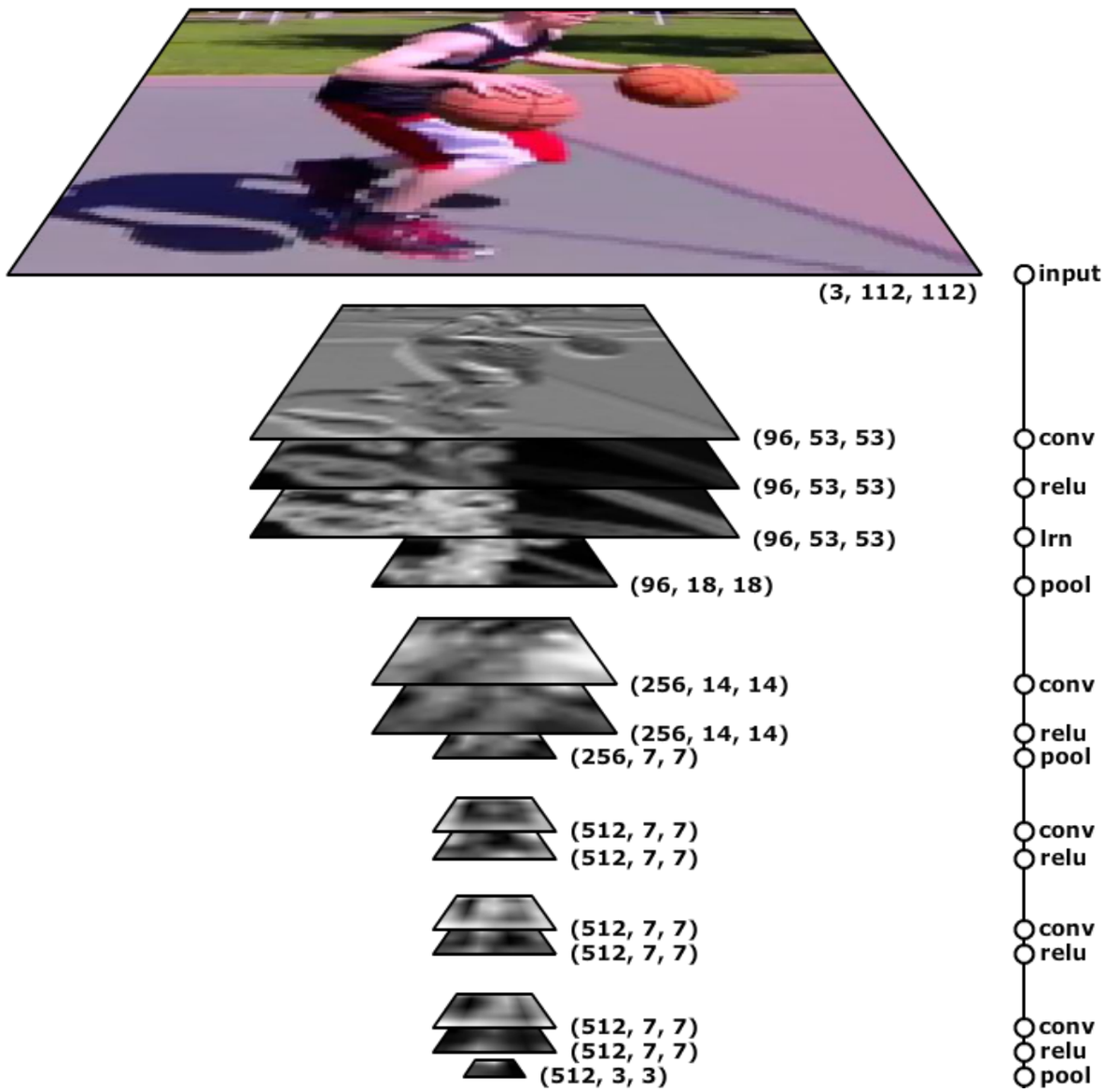


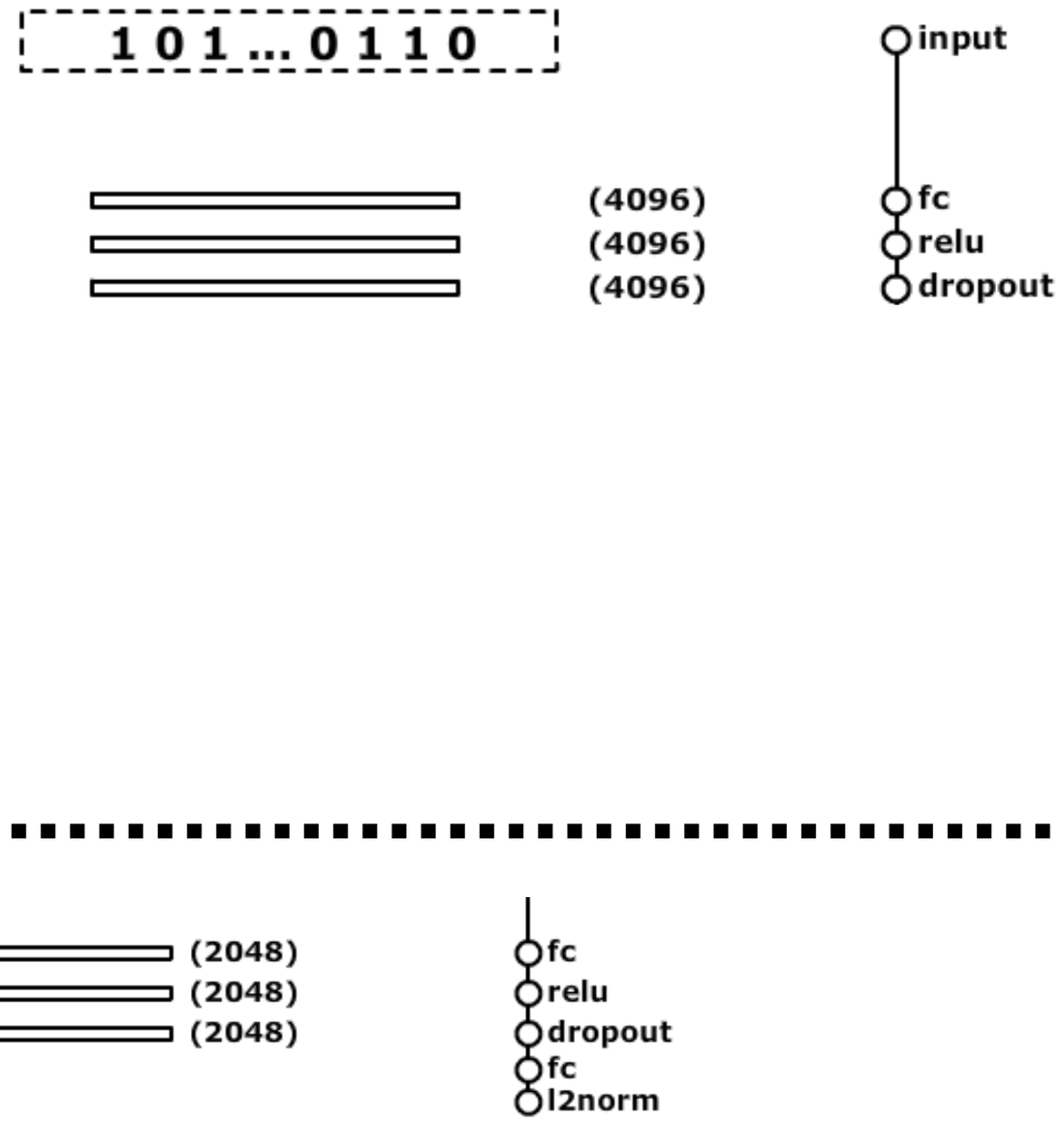
$$\mathcal{L}_c = 1 - f(e_{image}) \cdot g(e_{pose})$$

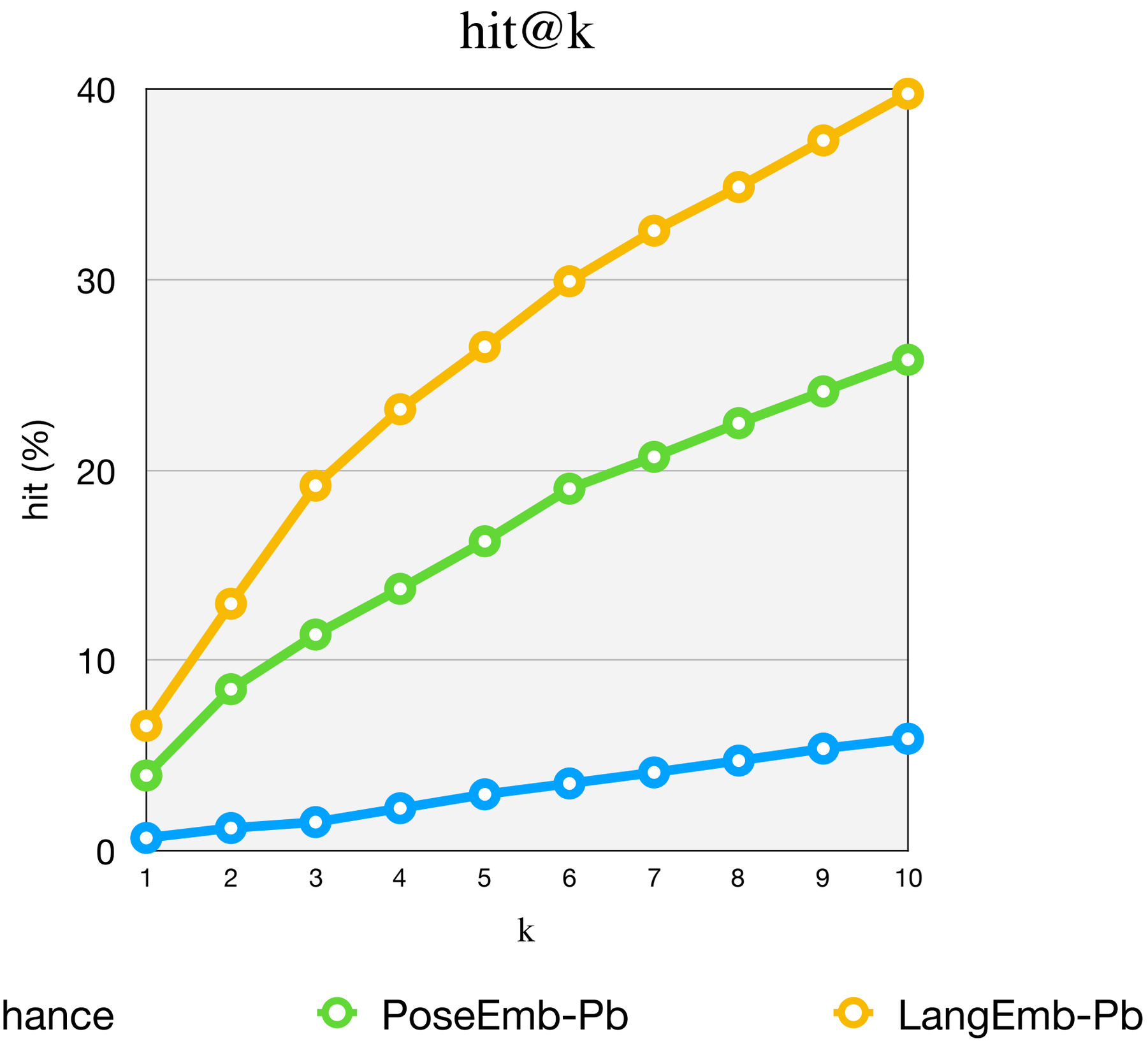

(3, 112, 112)
input
(96, 53, 53)
conv
(96, 53, 53)
relu
(96, 53, 53)
lrn
(96, 18, 18)
pool
(256, 14, 14)
conv
(256, 14, 14)
relu
(256, 7, 7)
pool
(512, 7, 7)
conv
(512, 7, 7)
relu
(512, 7, 7)
conv
(512, 7, 7)
relu
(512, 7, 7)
conv
(512, 7, 7)
relu
(512, 3, 3)
pool
(2048)
fc
(2048)
relu
(2048)
dropout
(128)
fc
(128)
l2norm

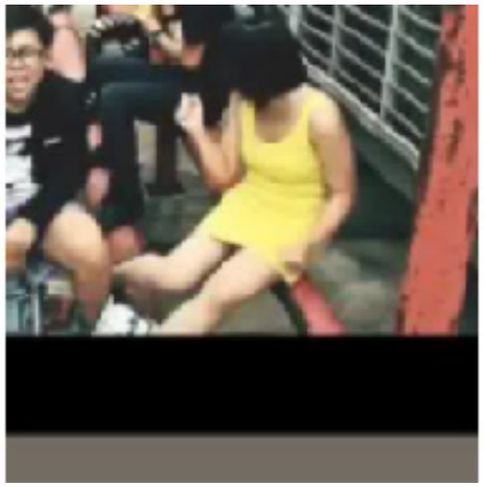

(2048)
(2048)
(2048)
(128)
(128)
fc
relu
dropout
fc
l2norm

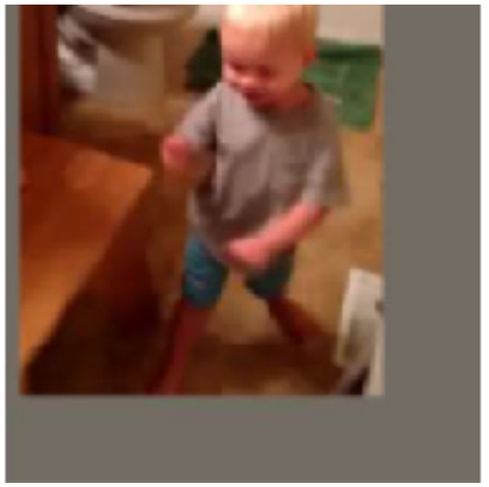

0 1 1 0 1 0 1 1 1 0 1
(4096)
(4096)
(4096)
(2048)
(2048)
(2048)
(128)
(128)
input
fc
relu
dropout
fc
relu
dropout
fc
l2norm


(3, 112, 112)
input
(96, 53, 53)
conv
(96, 53, 53)
relu
(96, 53, 53)
lrn
(96, 18, 18)
pool
(256, 14, 14)
conv
(256, 14, 14)
relu
(256, 7, 7)
pool
(512, 7, 7)
conv
(512, 7, 7)
relu
(512, 7, 7)
conv
(512, 7, 7)
relu
(512, 7, 7)
conv
(512, 7, 7)
relu
(512, 3, 3)
pool

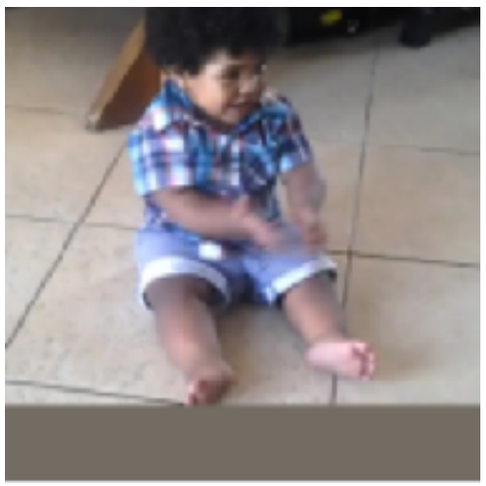

1 0 1 ... 0 1 1 0
(4096)
(4096)
(4096)
input
fc
relu
dropout
(2048)
(2048)
(2048)
(128)
(128)
fc
relu
dropout
fc
l2norm


hit@k
hit (%)
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
k
Chance
PoseEmb-Pb
LangEmb-Pb





1
0
0
⋮
1
1



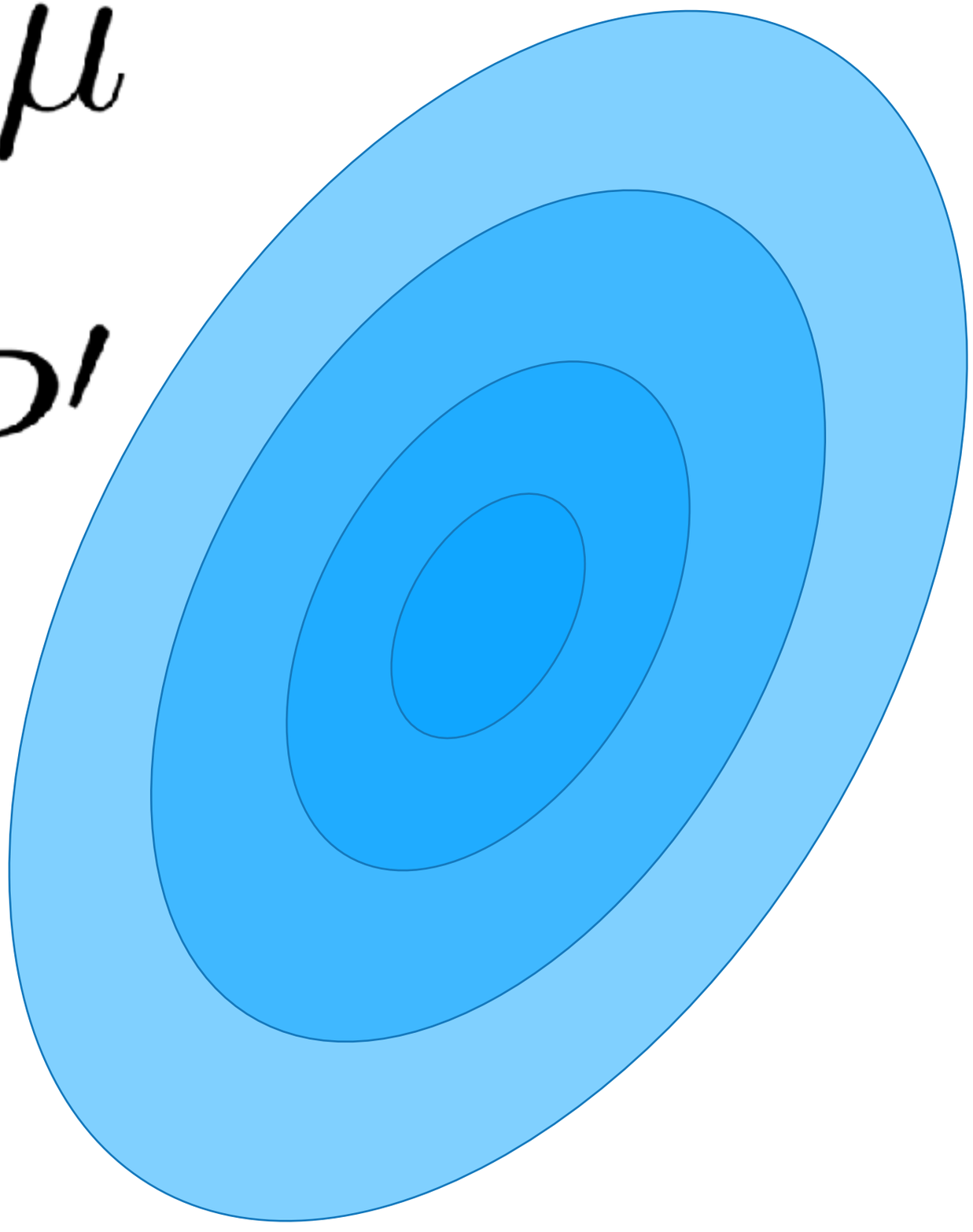
Language Subset Querying

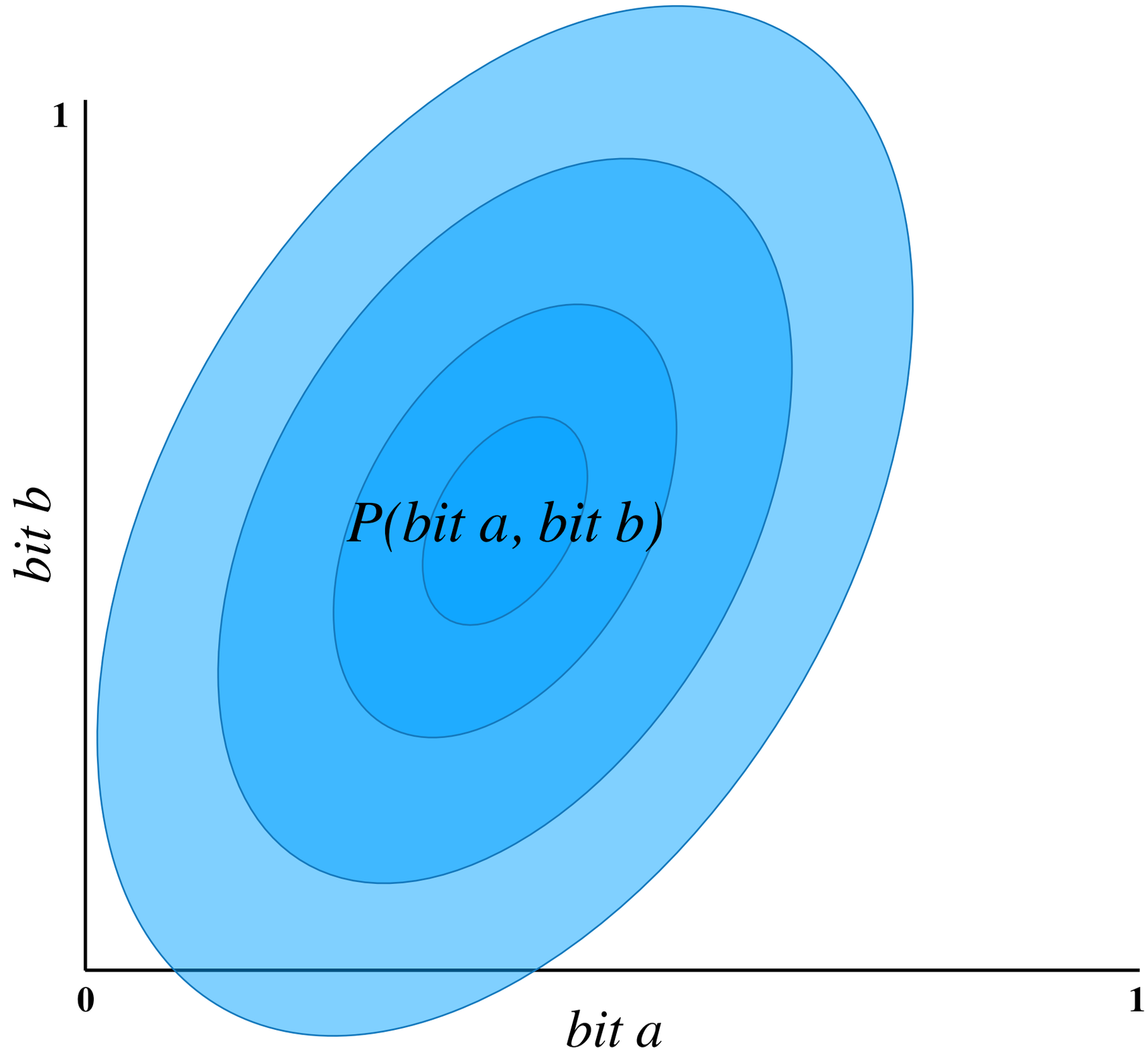

1 1

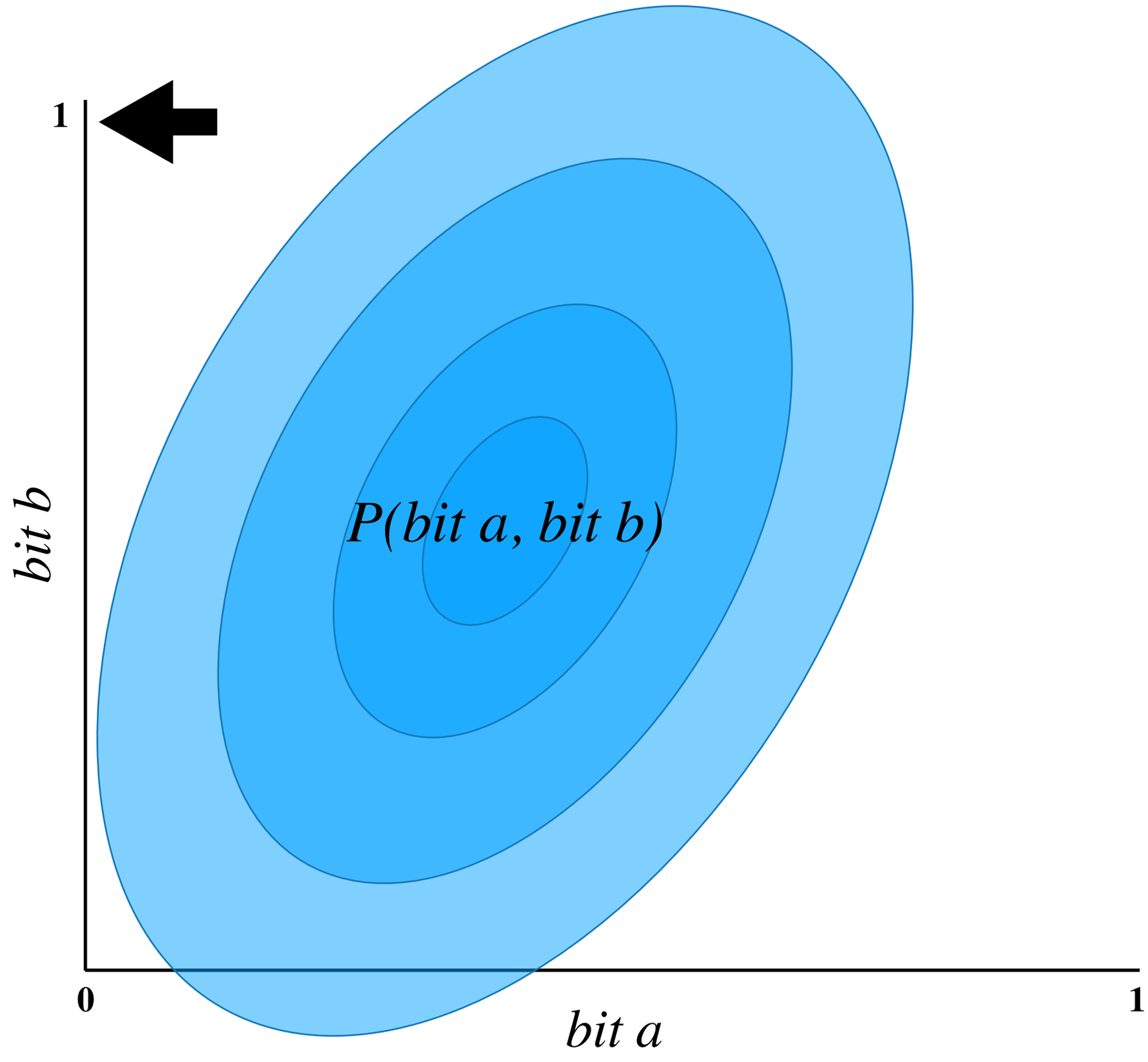
?	1	1	?	?	?	?	?	?	?	?

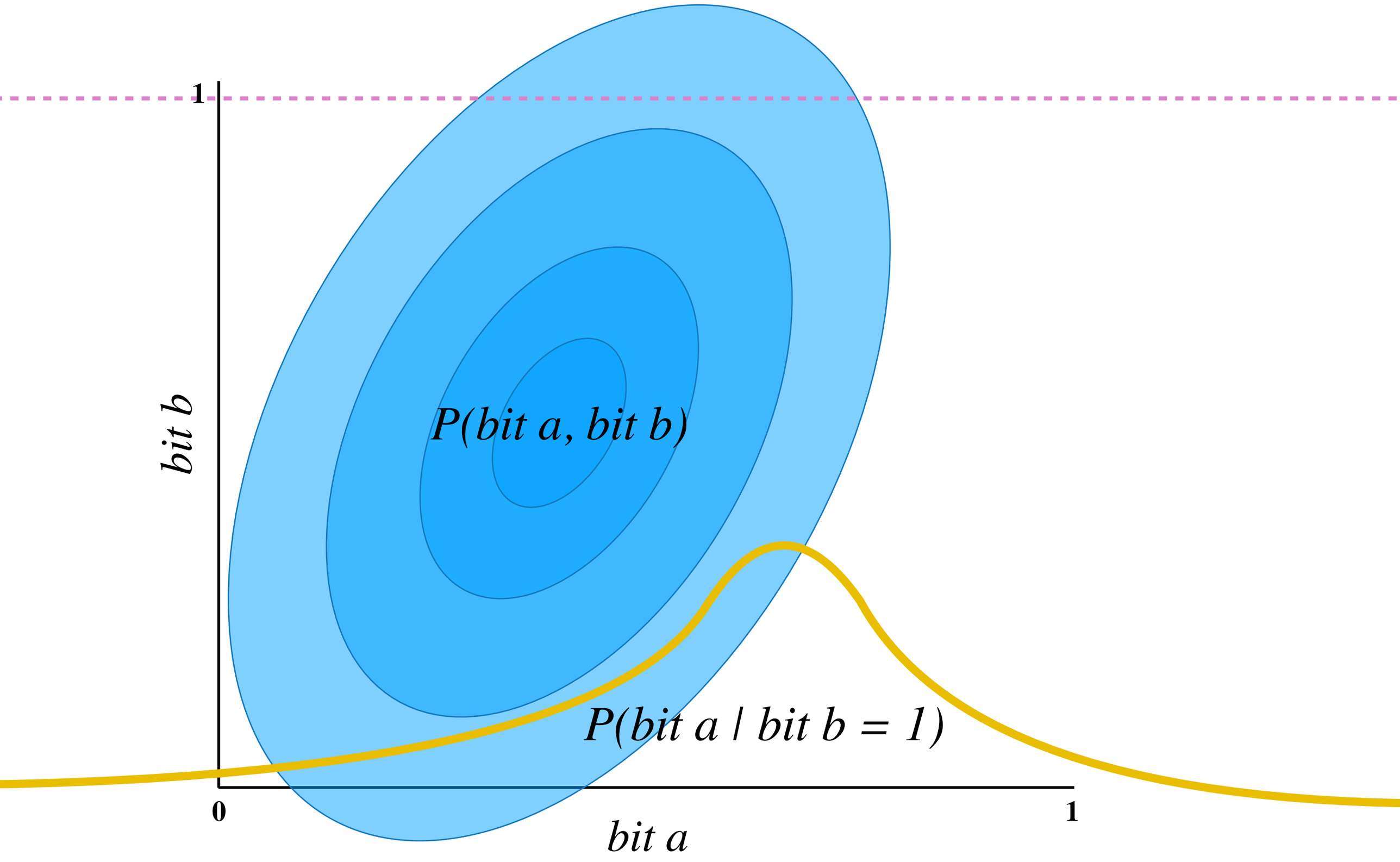
$$P' = P - \mu$$

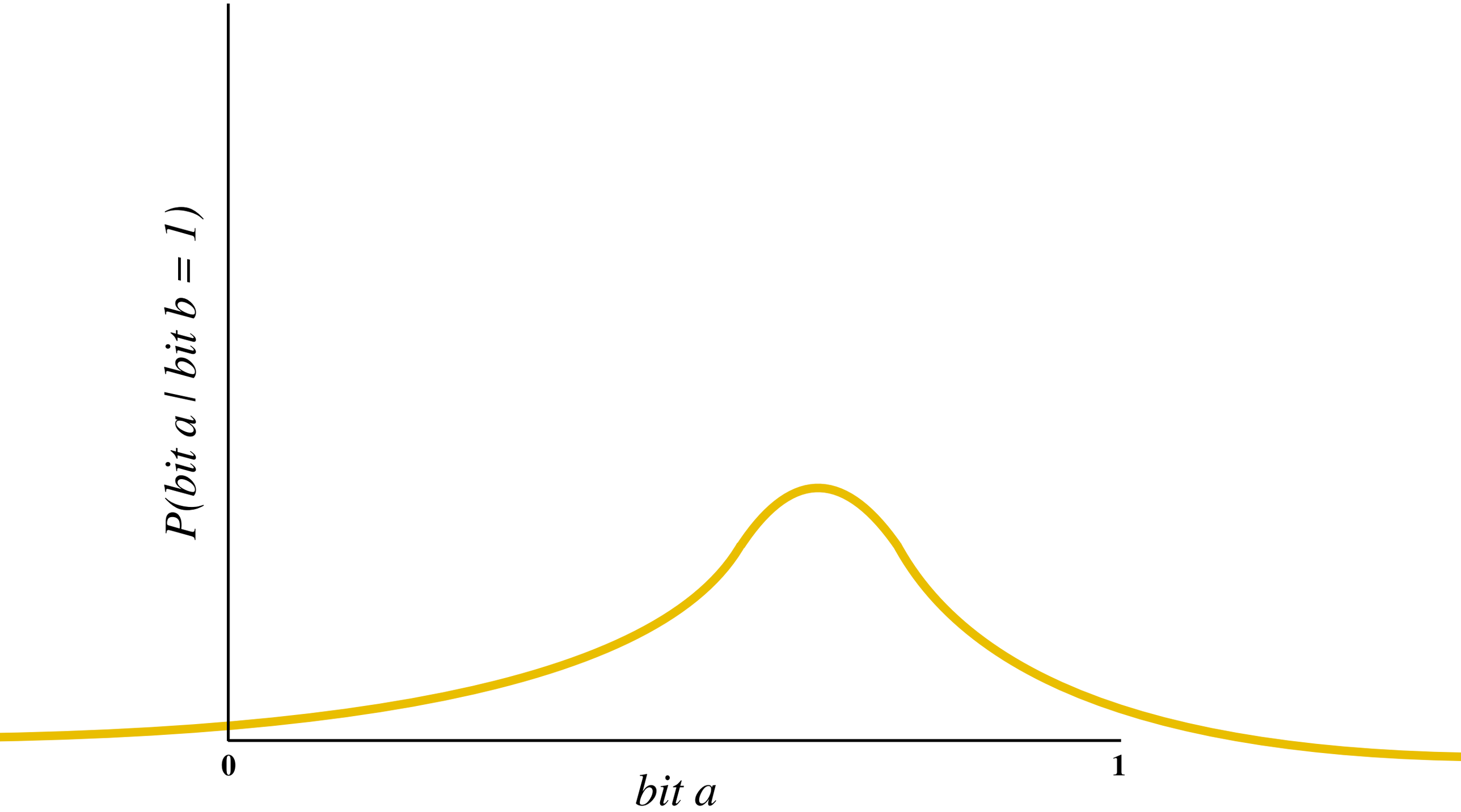
$$\sum = P'^T P'$$

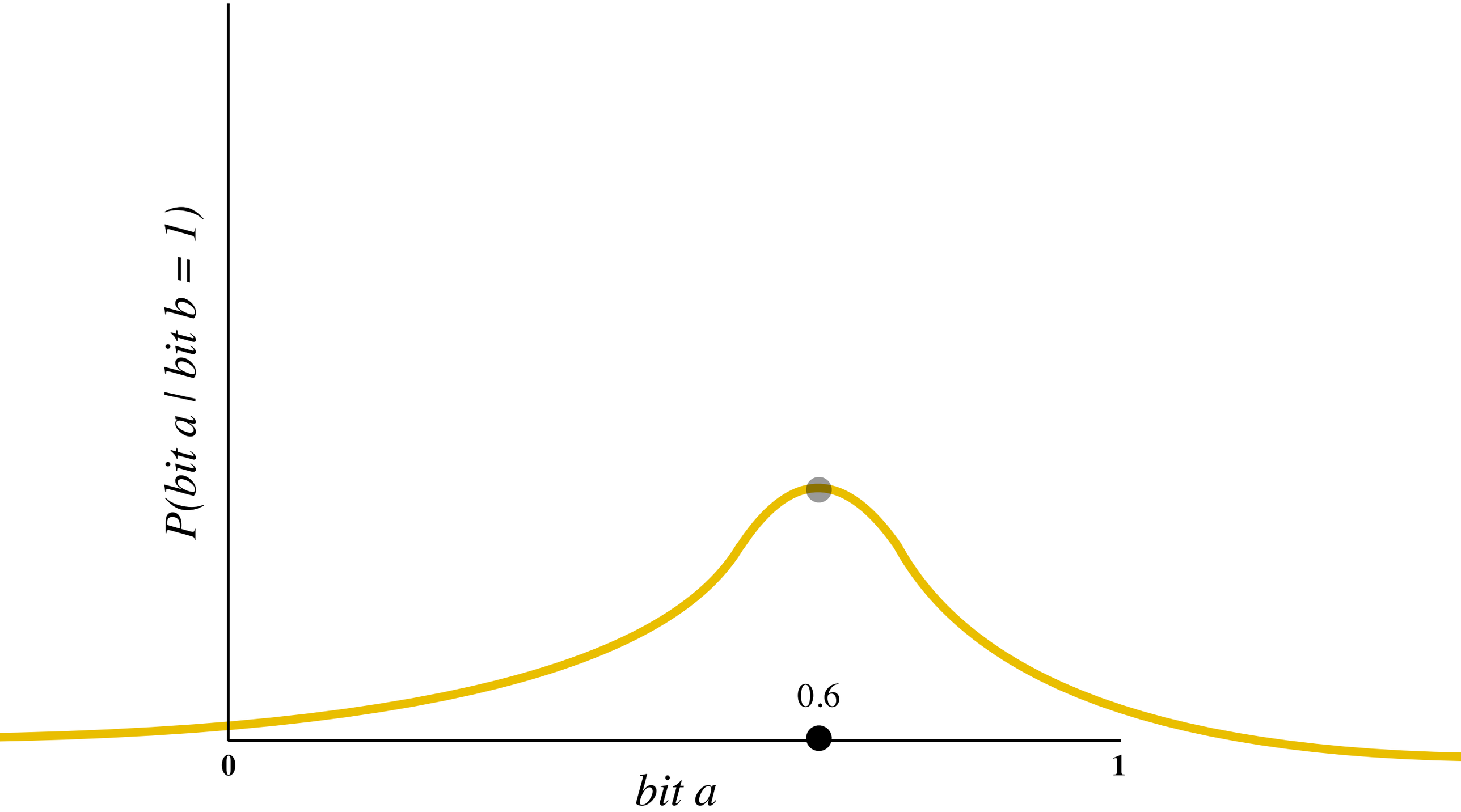



1
bit b
P(bit a, bit b)
0
1
bit a


1
bit b
P(bit a, bit b)
0
1
bit a

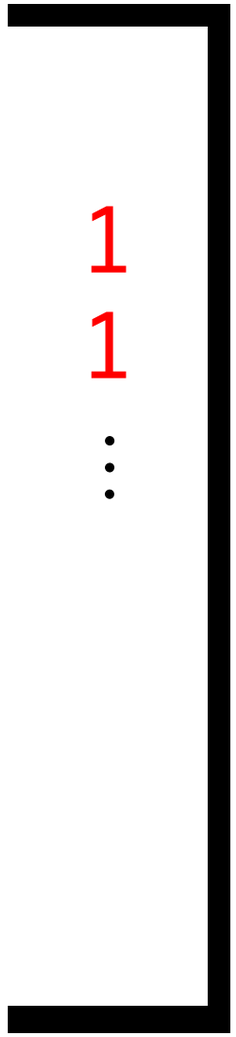

1
bit b
P(bit a, bit b)
P(bit a | bit b = 1)
0
1
bit a

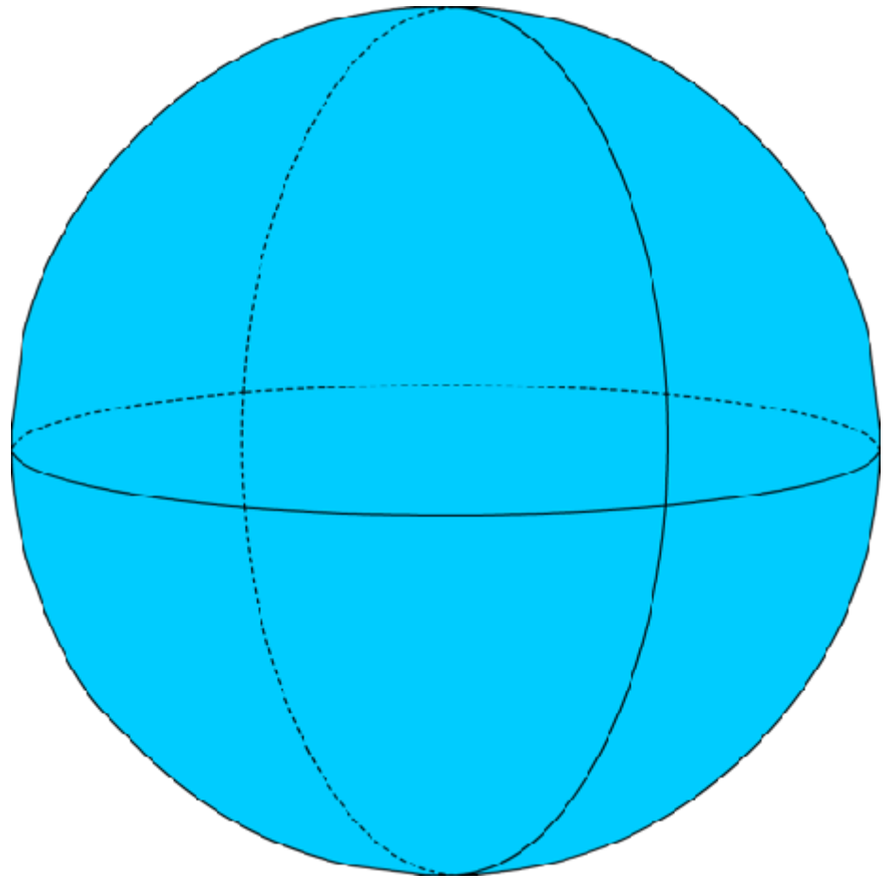

P(bit a / bit b = 1)
0
1
bit a

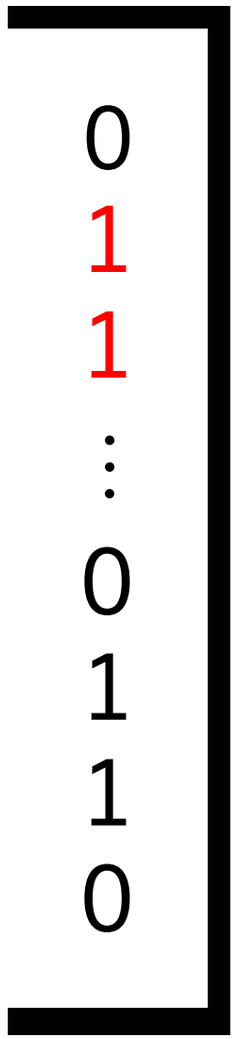

P(bit a / bit b = 1)
0.6
0
1
bit a

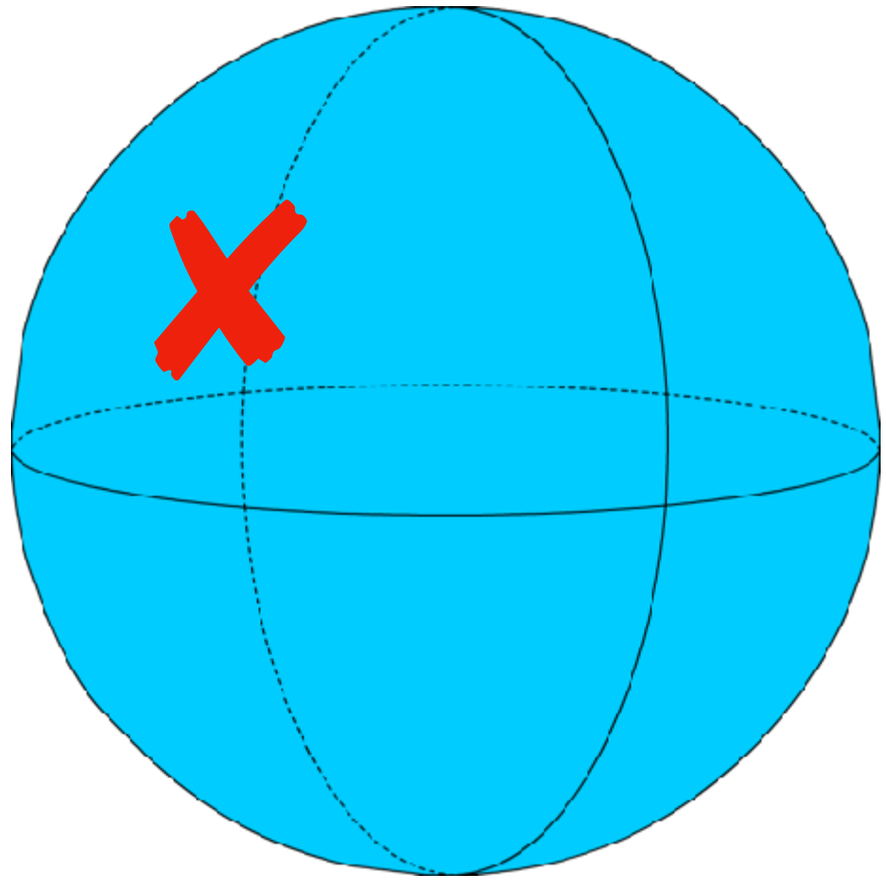

1 1

0	1	1	1	0	1	1	1	0	1	1


1
1
⋮

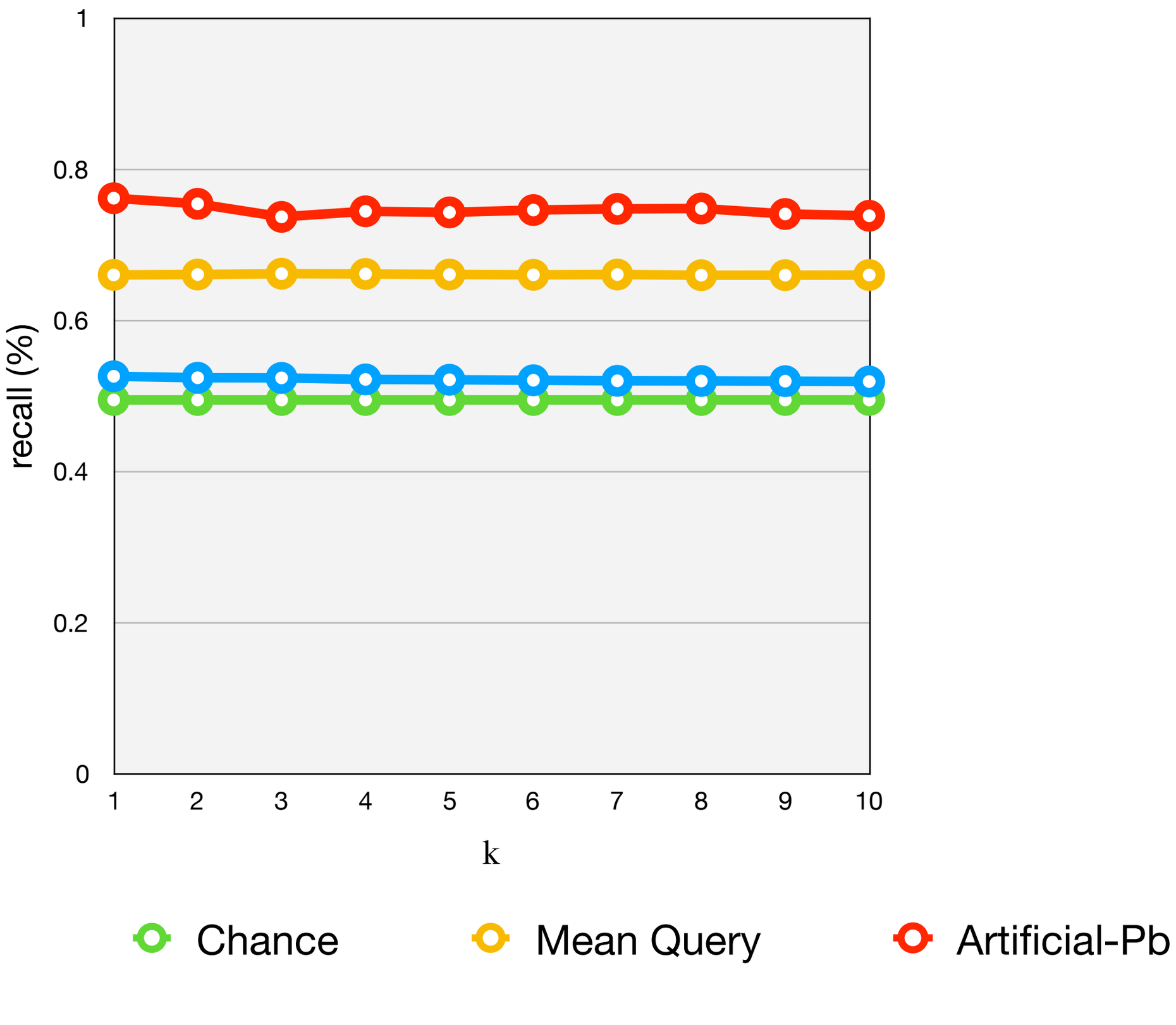


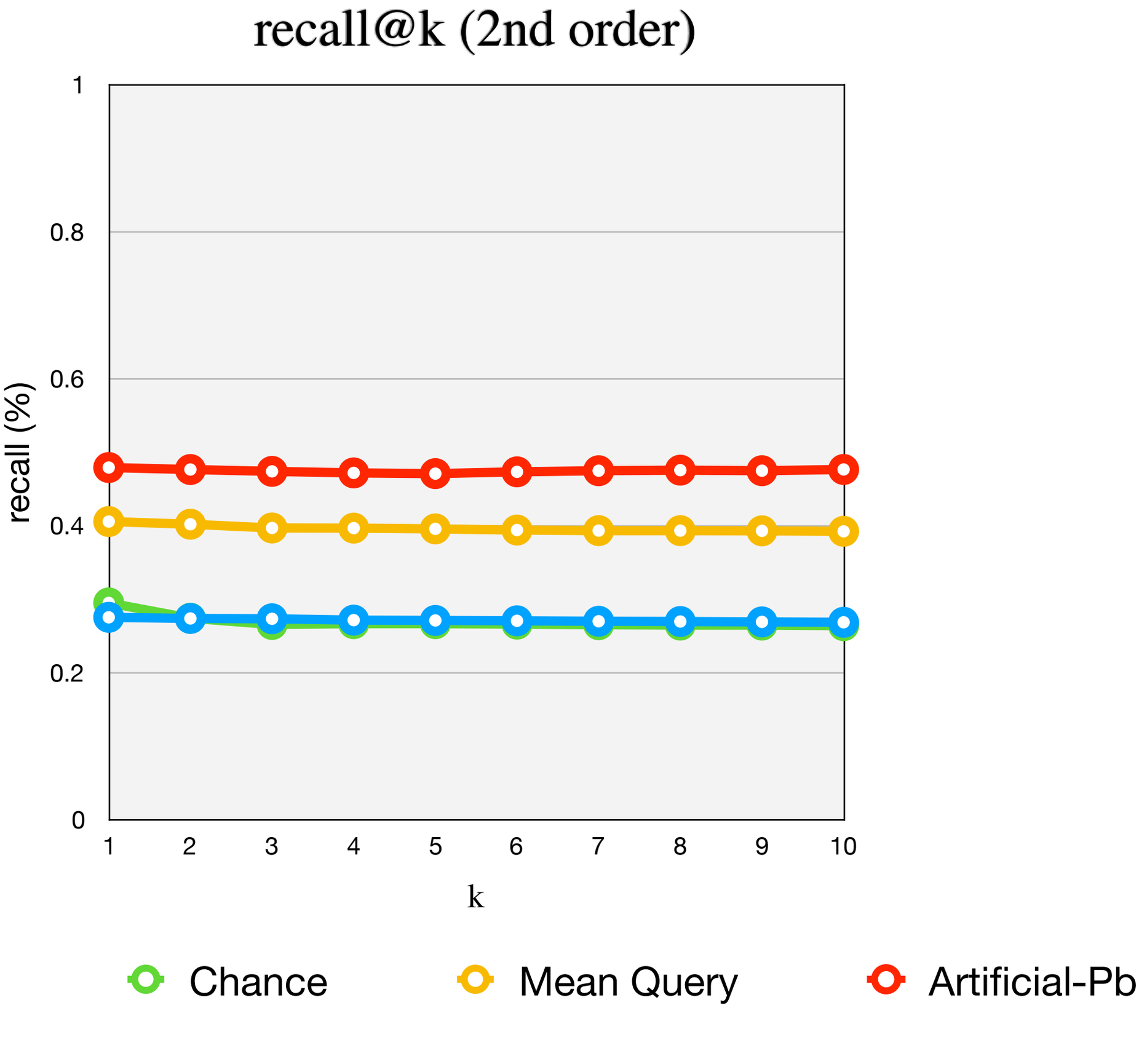

0
1
1
…
0
1
1
0



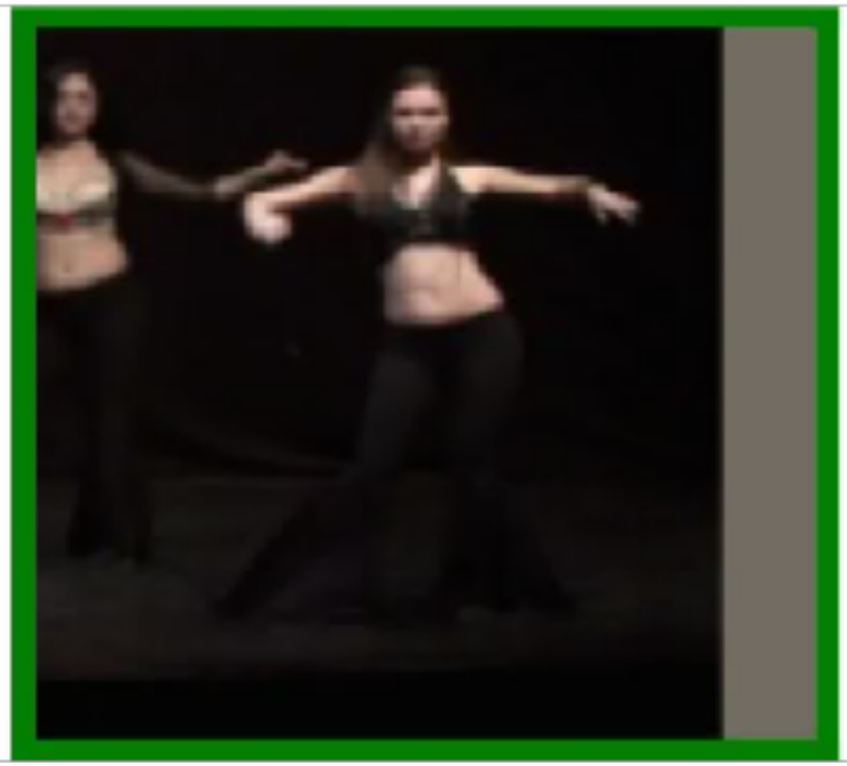
$$recall@k = \frac{1}{N} \sum_{i=1}^{N} recall_k(R_{1...k})$$

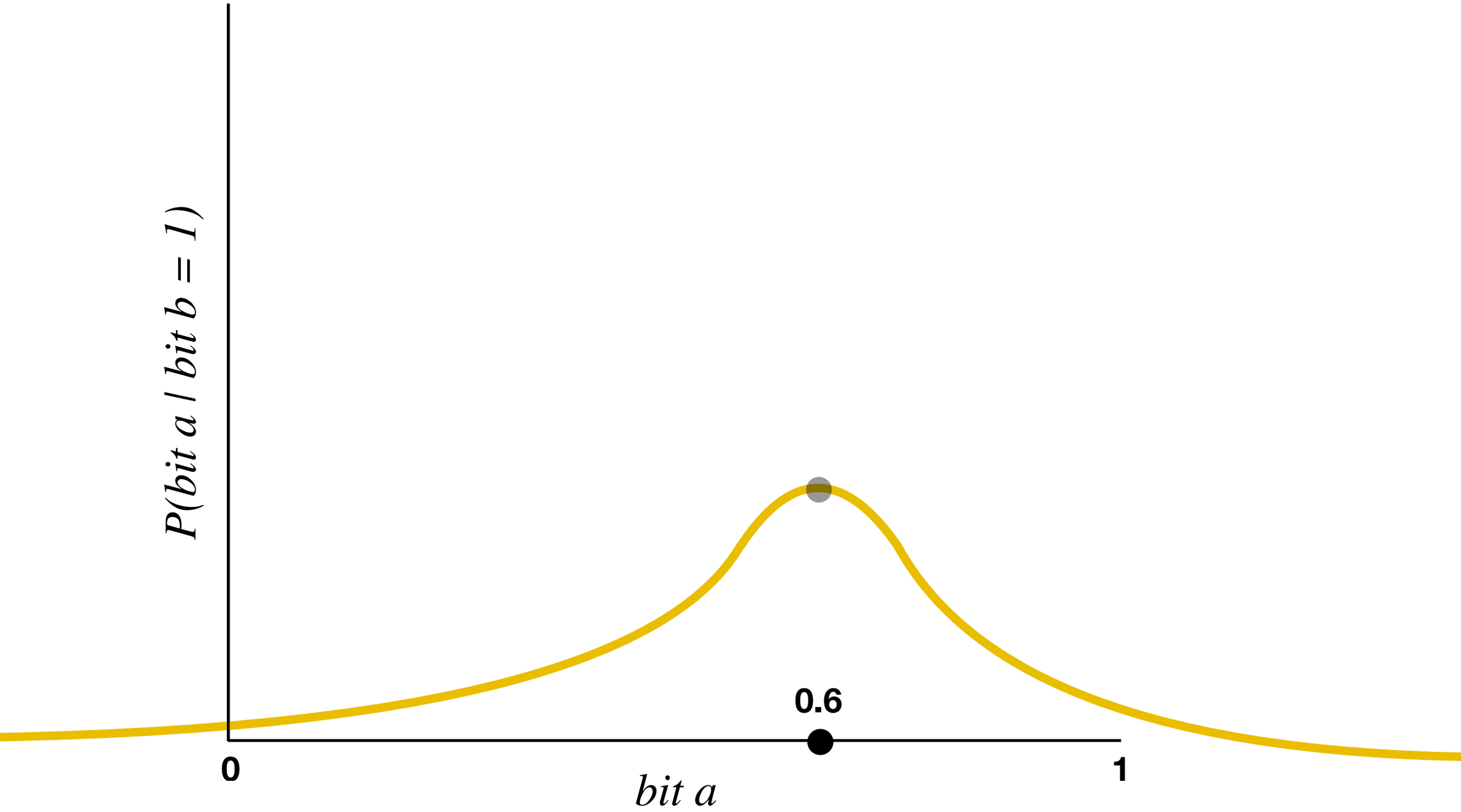
$$recall_k(X) = \frac{|X \cap X_{tp}|}{|X|}$$

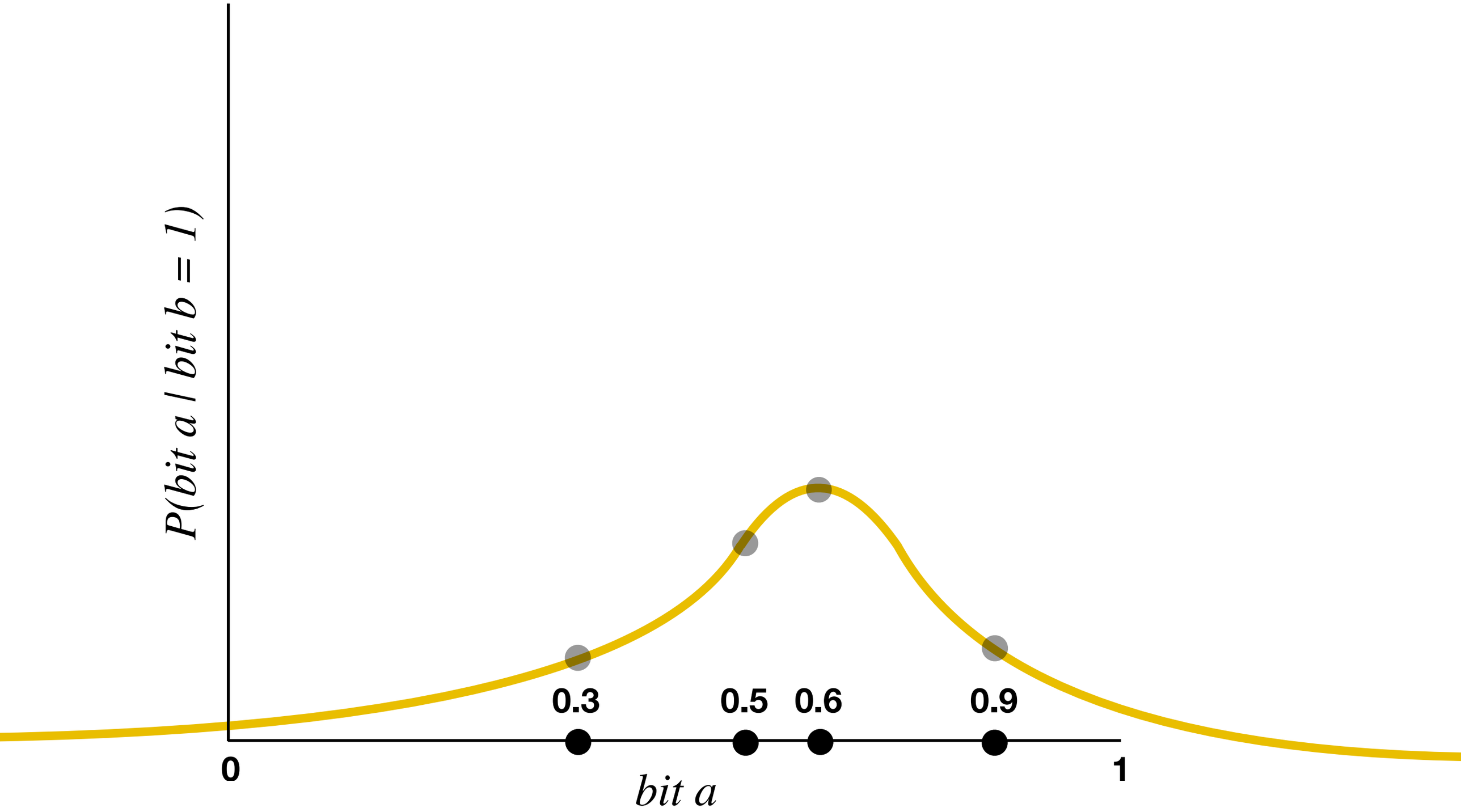

recall@k (1st order)
1
0.8
0.6
0.4
0.2
0
recall (%)
1
2
3
4
5
6
7
8
9
10
k
VGG-S
Chance
Mean Query
Artificial-Pb

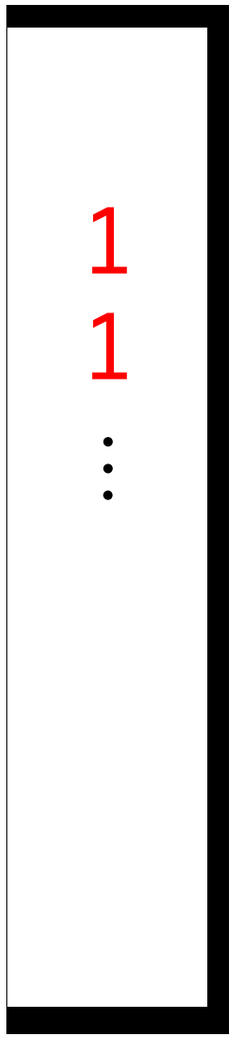

recall@k (2nd order)
recall (%)
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
7
8
9
10
k
VGG-S
Chance
Mean Query
Artificial-Pb

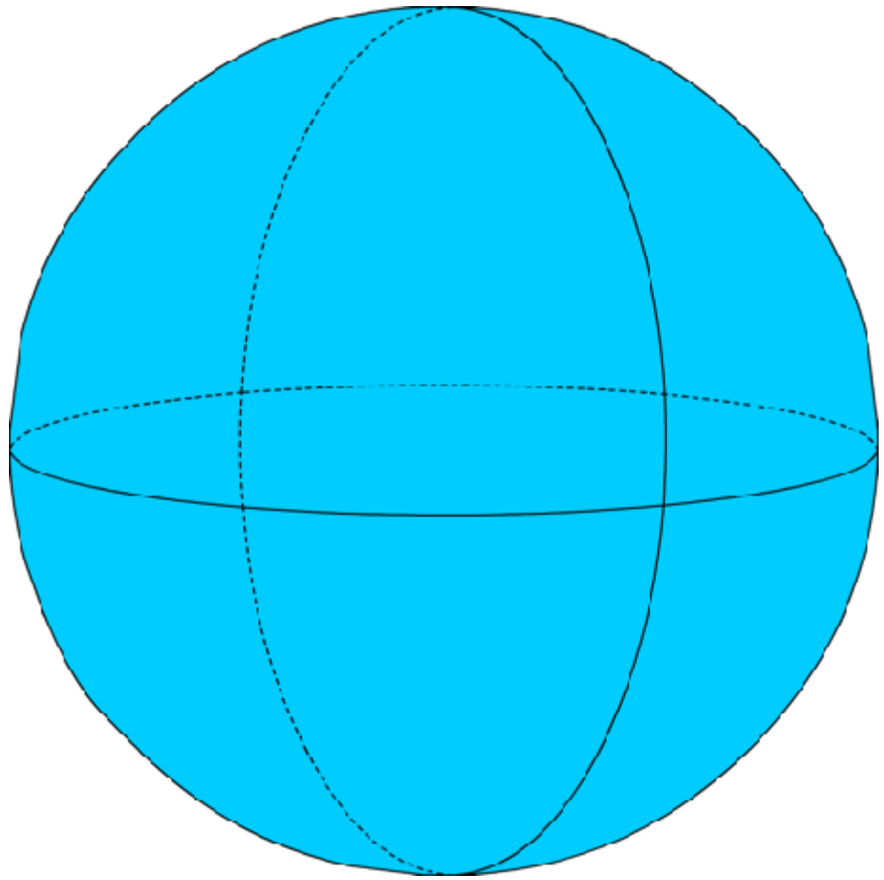
right shoulder **is bent**

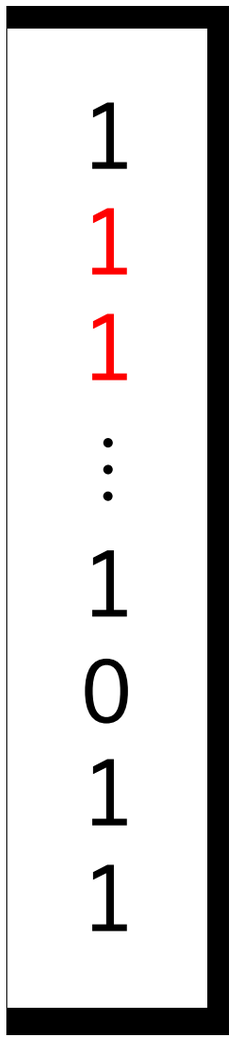


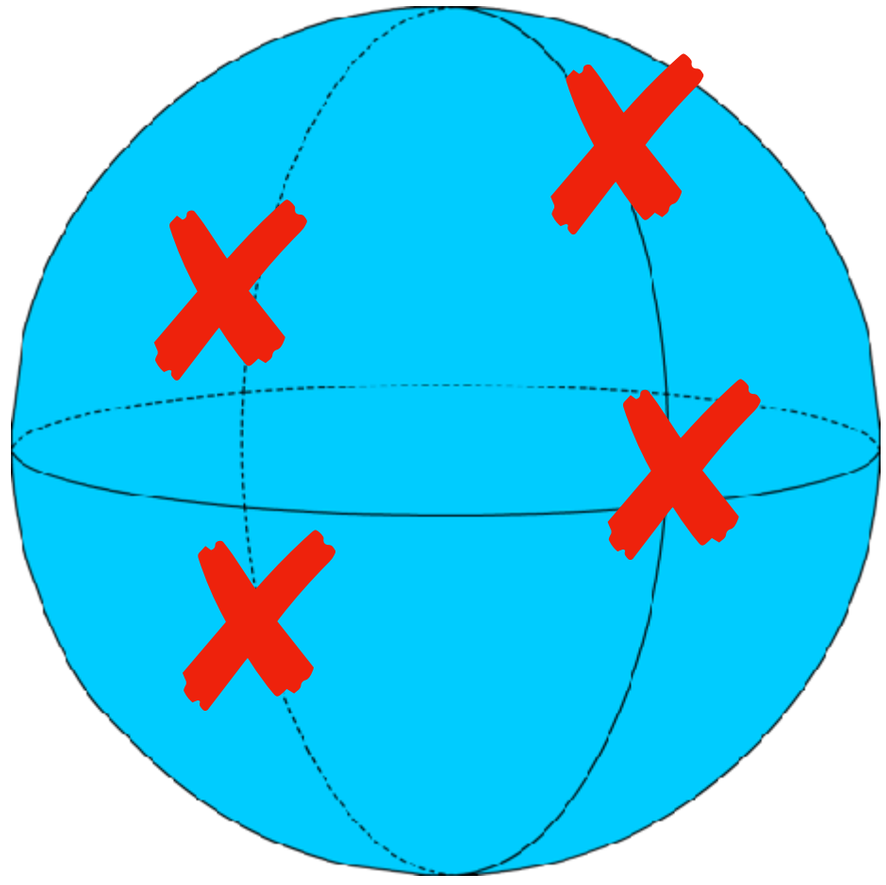

P(bit a / bit b = 1)
0.6
0
1
bit a


P(bit a / bit b = 1)
0.3
0.5
0.6
0.9
0
1
bit a


1
1
⋮



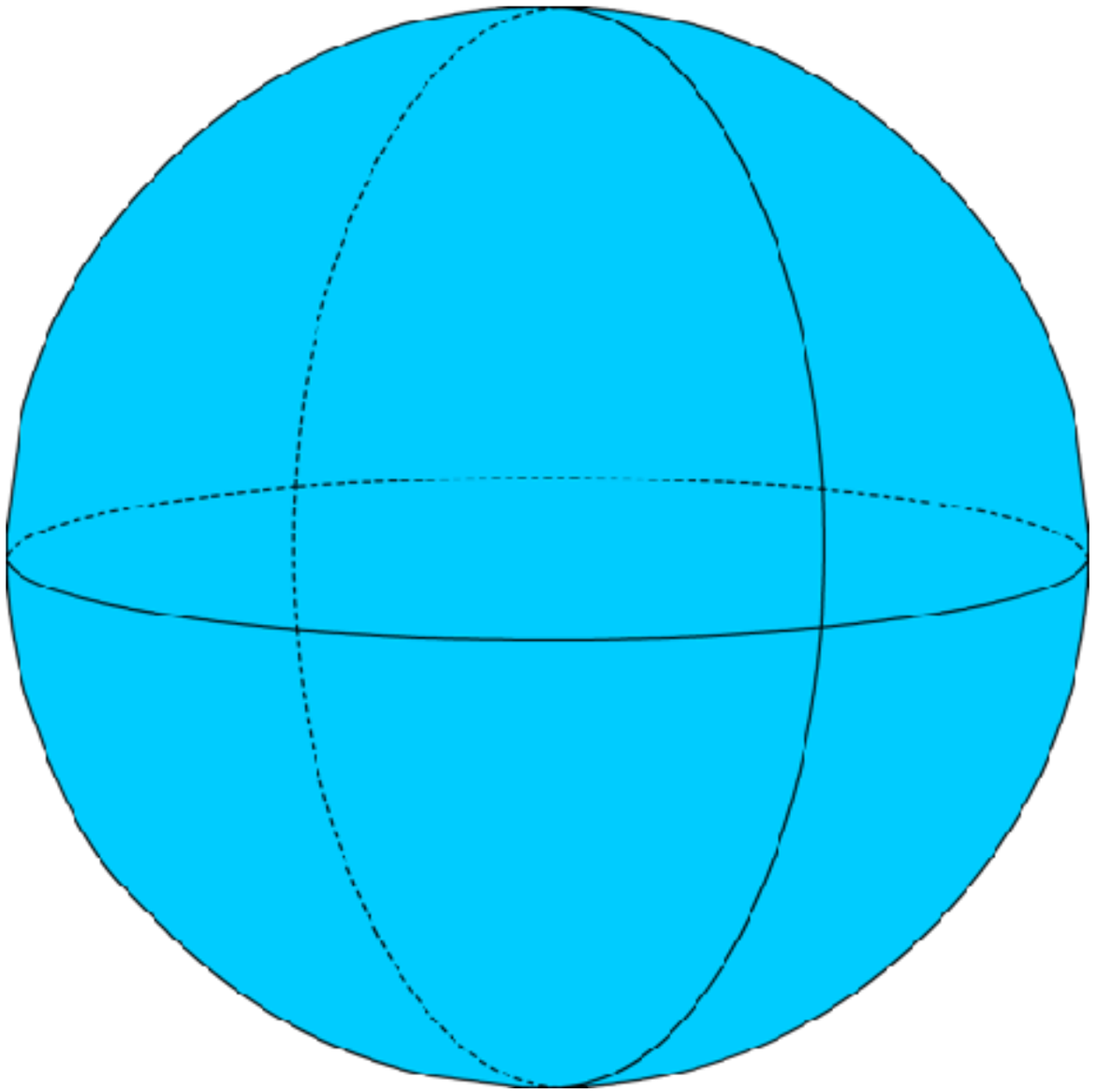

1
1
1
1
0
1
1



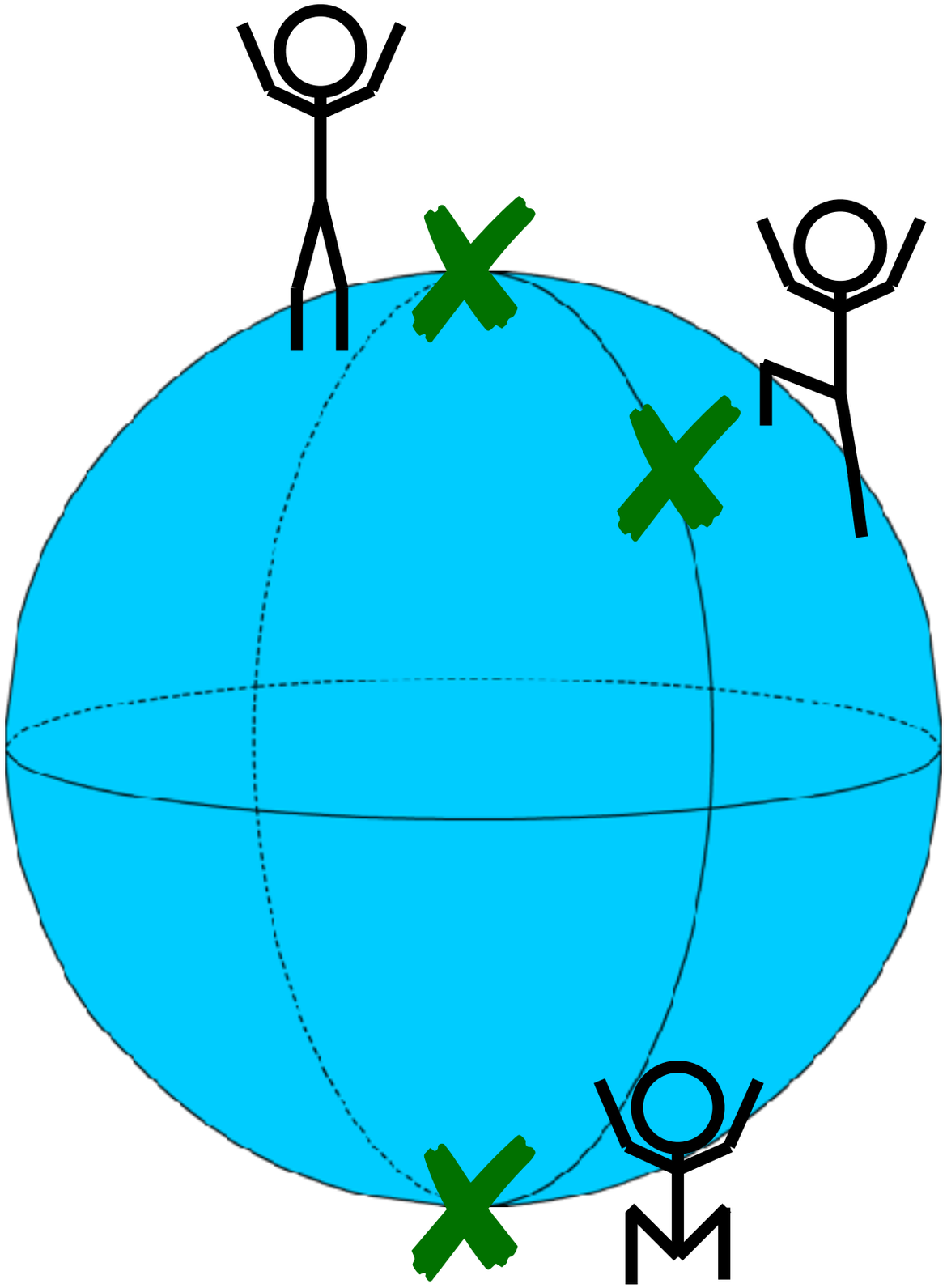
right shoulder **is bent**

Mahalanobis Distance	mean recall@1
0.0	76.2%
0.1	73.1%
1.0	71.1%
10.0	64.7%
Baseline	
Mean Query	66.1%

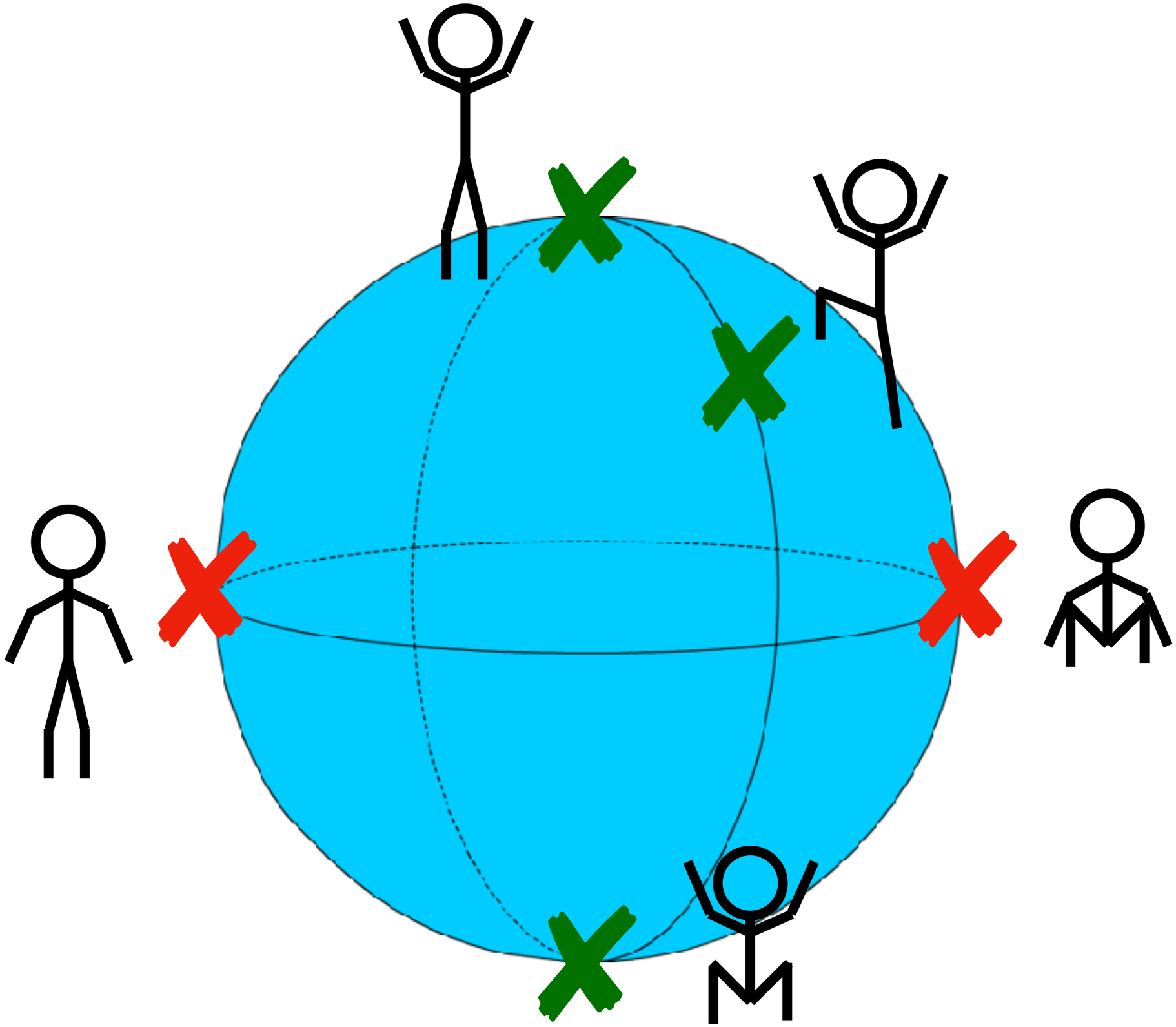
“Left hand above head; right hand above head.”



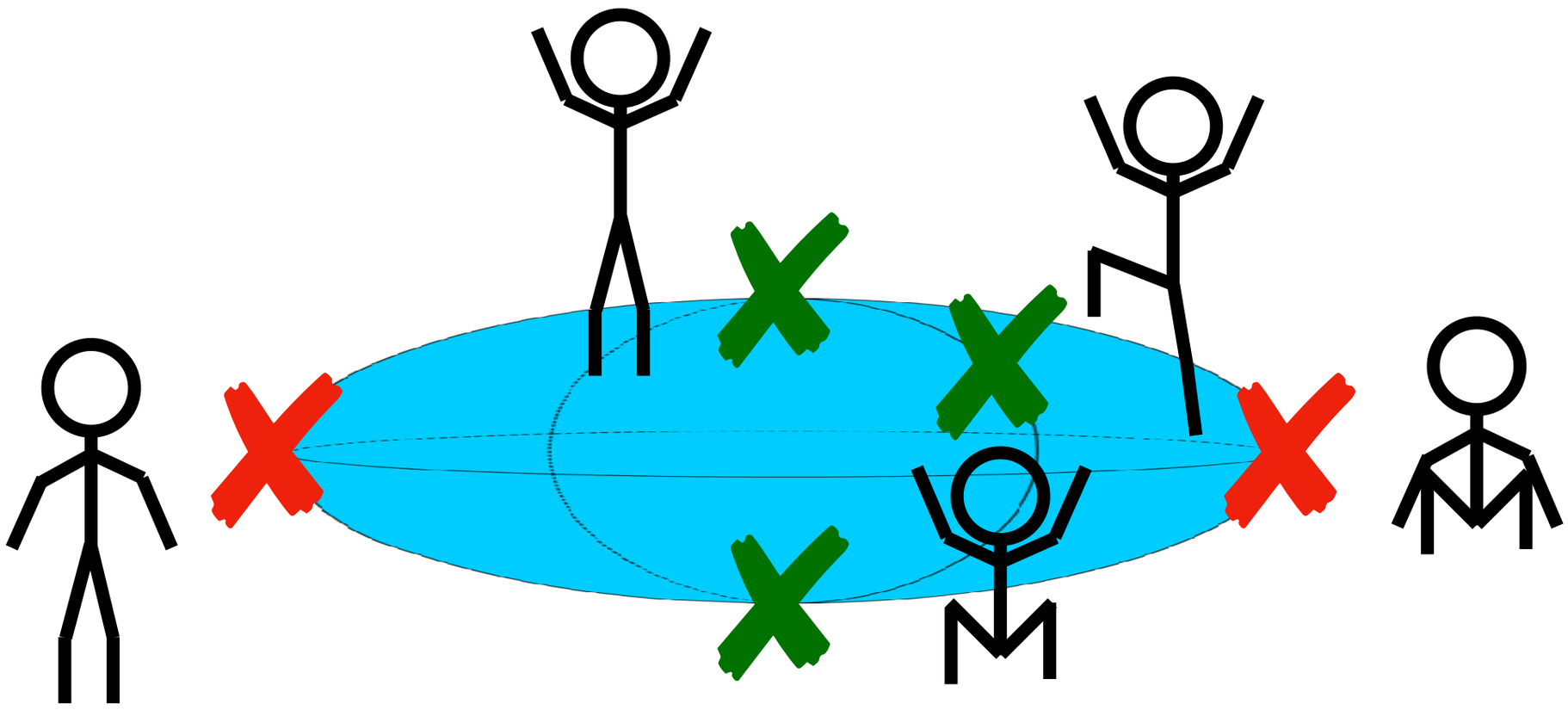
“Left hand above head; right hand above head.”



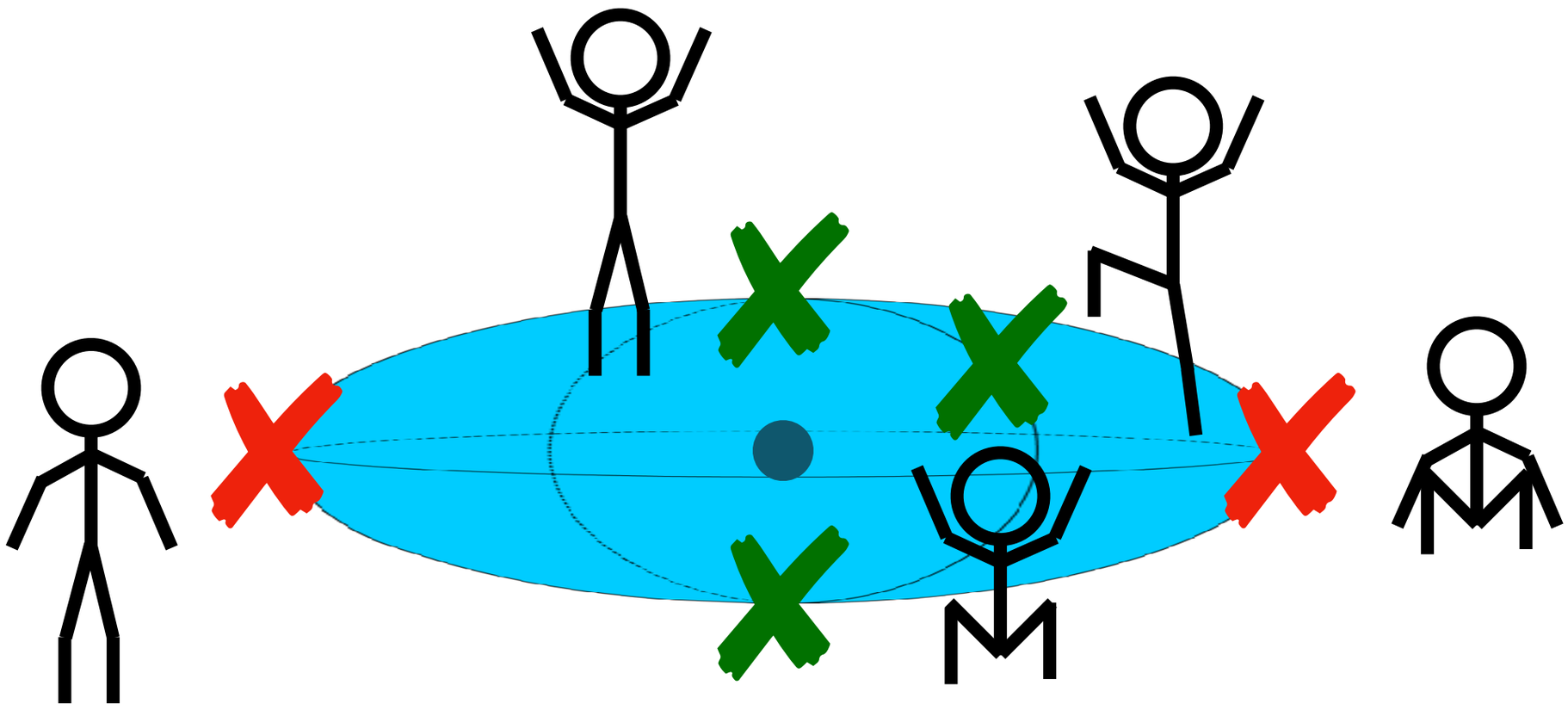
“Left hand above head; right hand above head.”



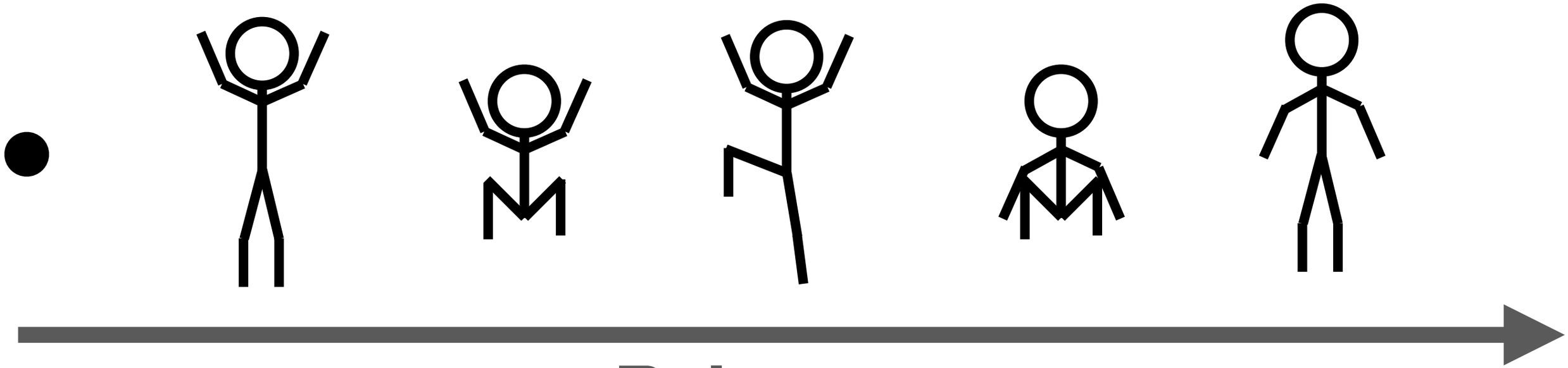
“Left hand above head; right hand above head.”



“Left hand above head; right hand above head.”

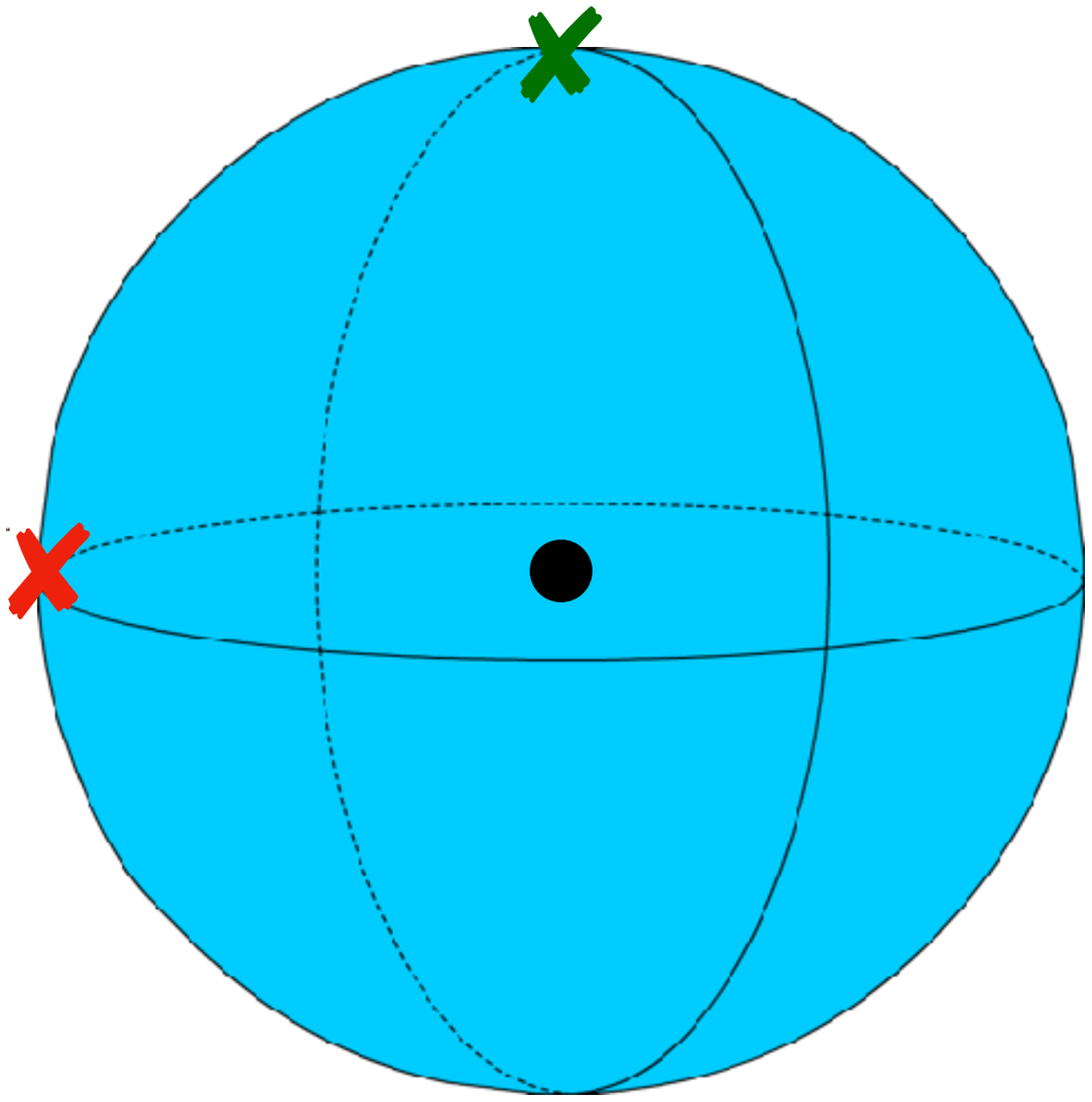


“Left hand above head; right hand above head.”

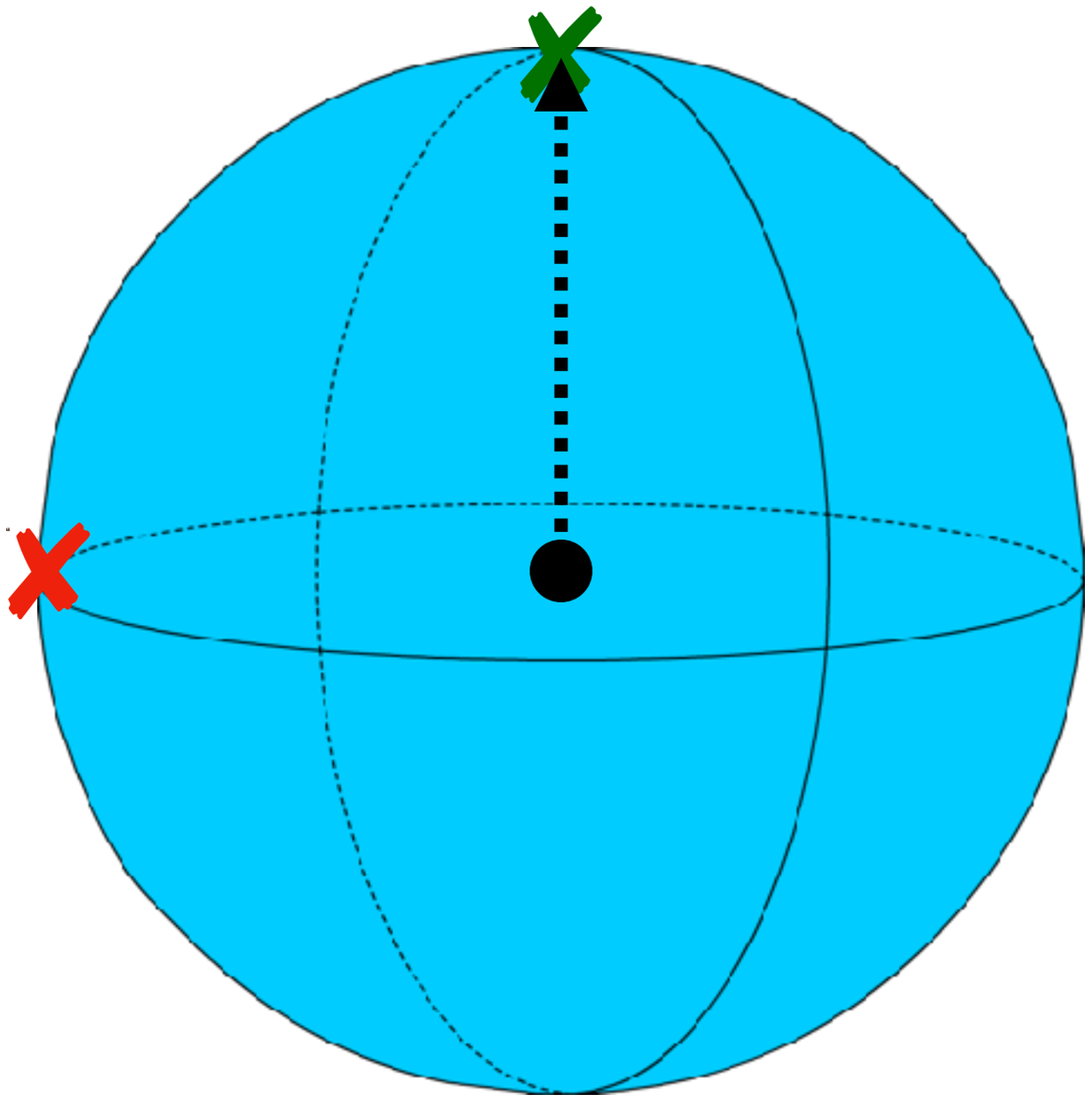


Relevance

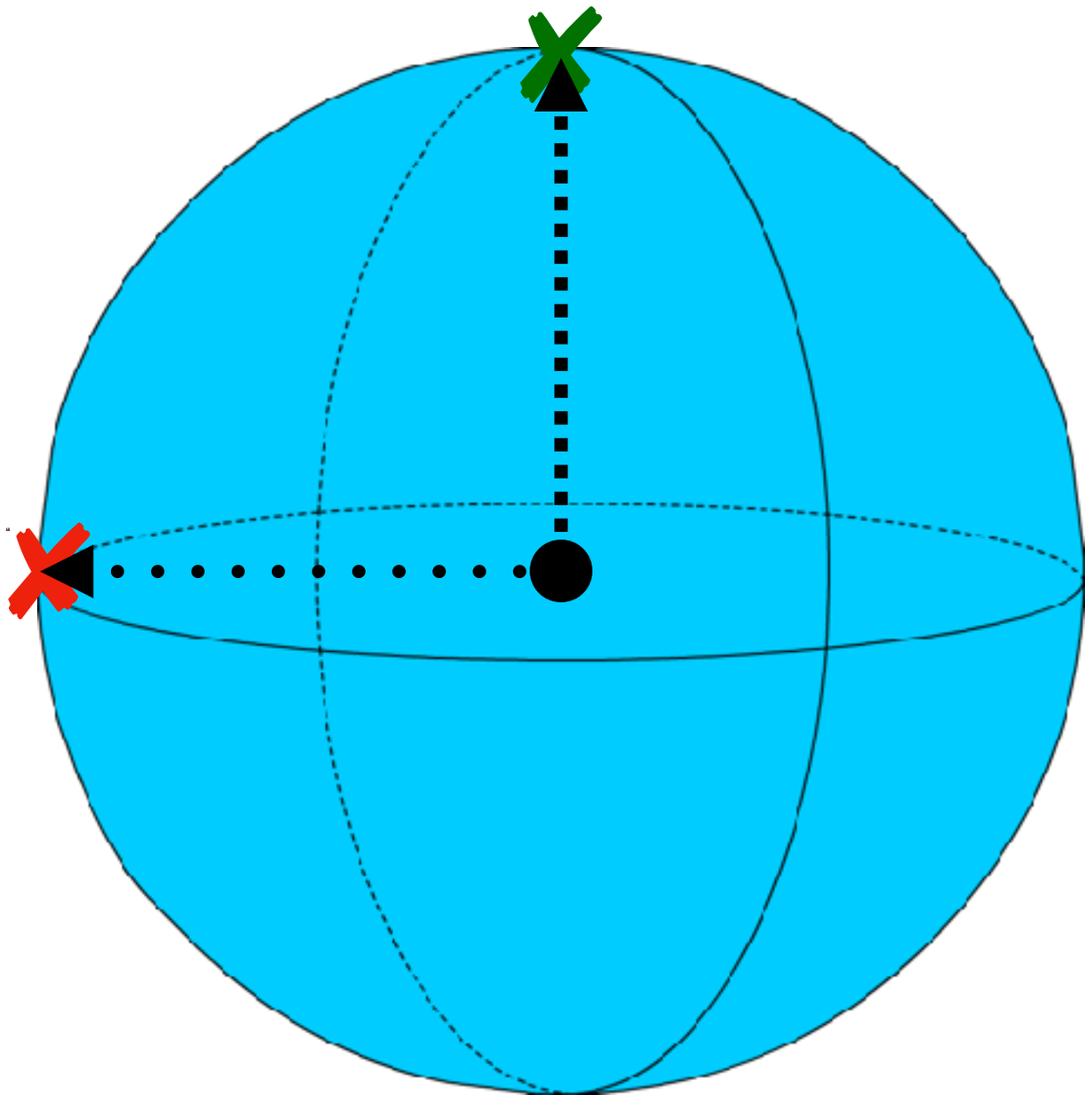
$$\left\{ a \; : \; p \; , \; n \right\}$$



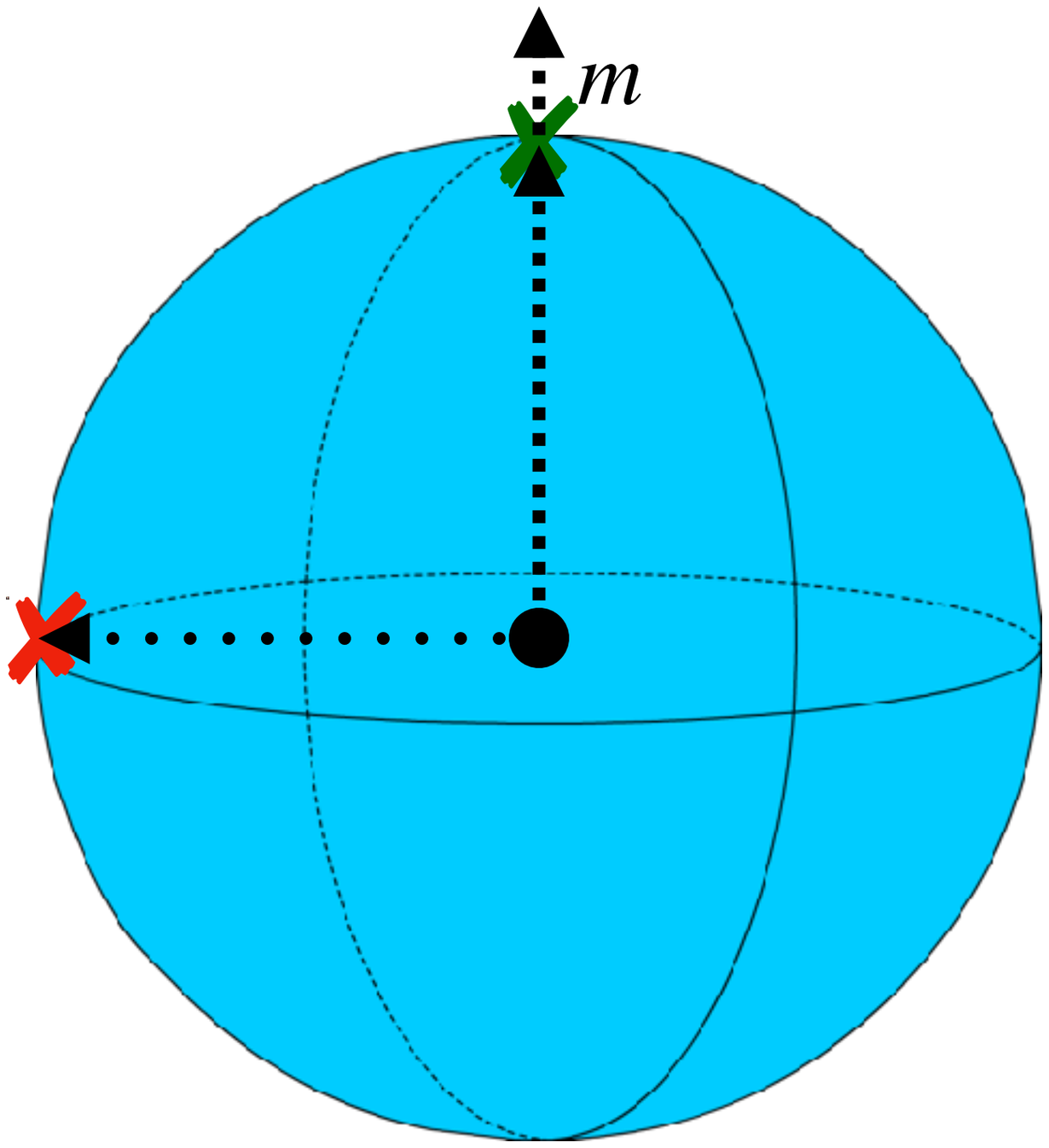
$$\mathcal{L}_{triplet} = \left[||f(a) - f(p)||_2^2 - ||f(a) - f(n)||_2^2 + m\right]_+$$



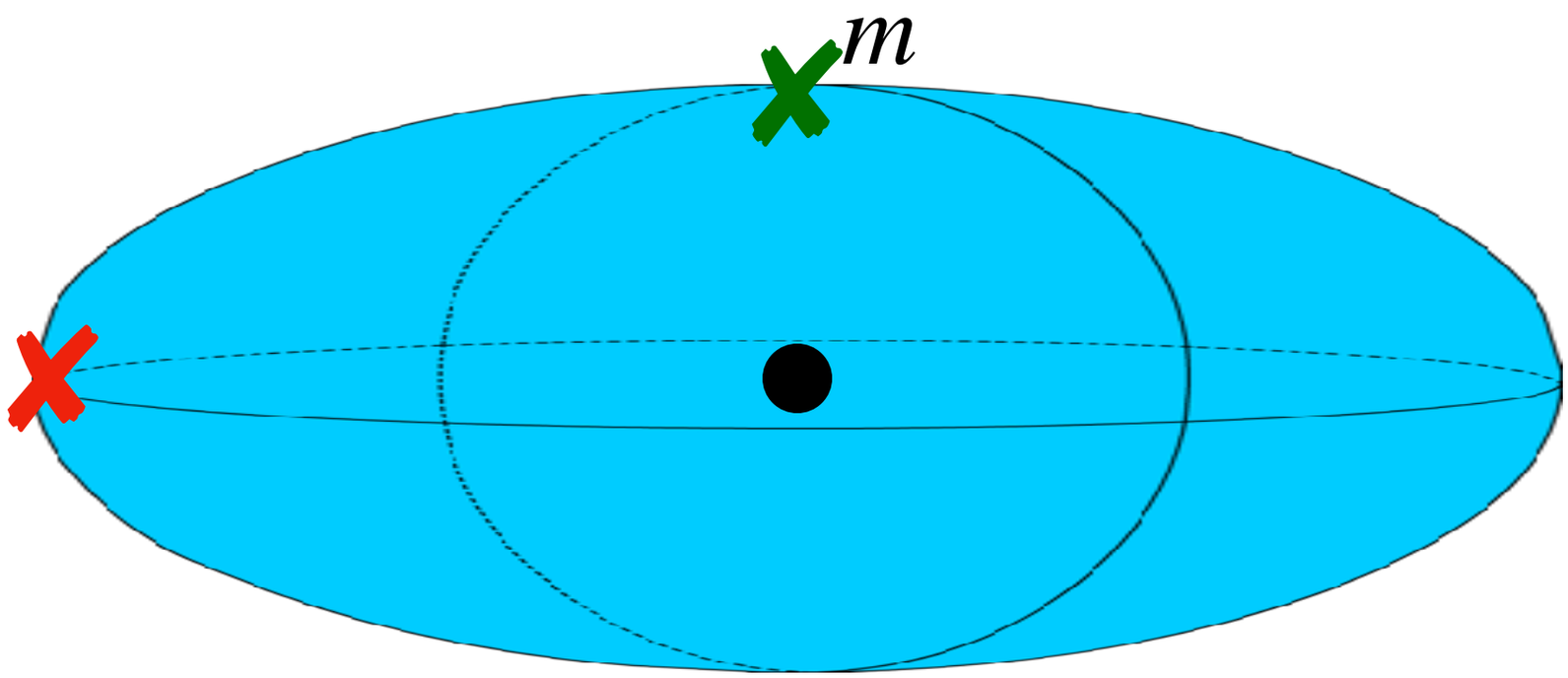
$$\mathcal{L}_{triplet} = \left[||f(a) - f(p)||_2^2 - ||f(a) - f(n)||_2^2 + m\right]_+$$



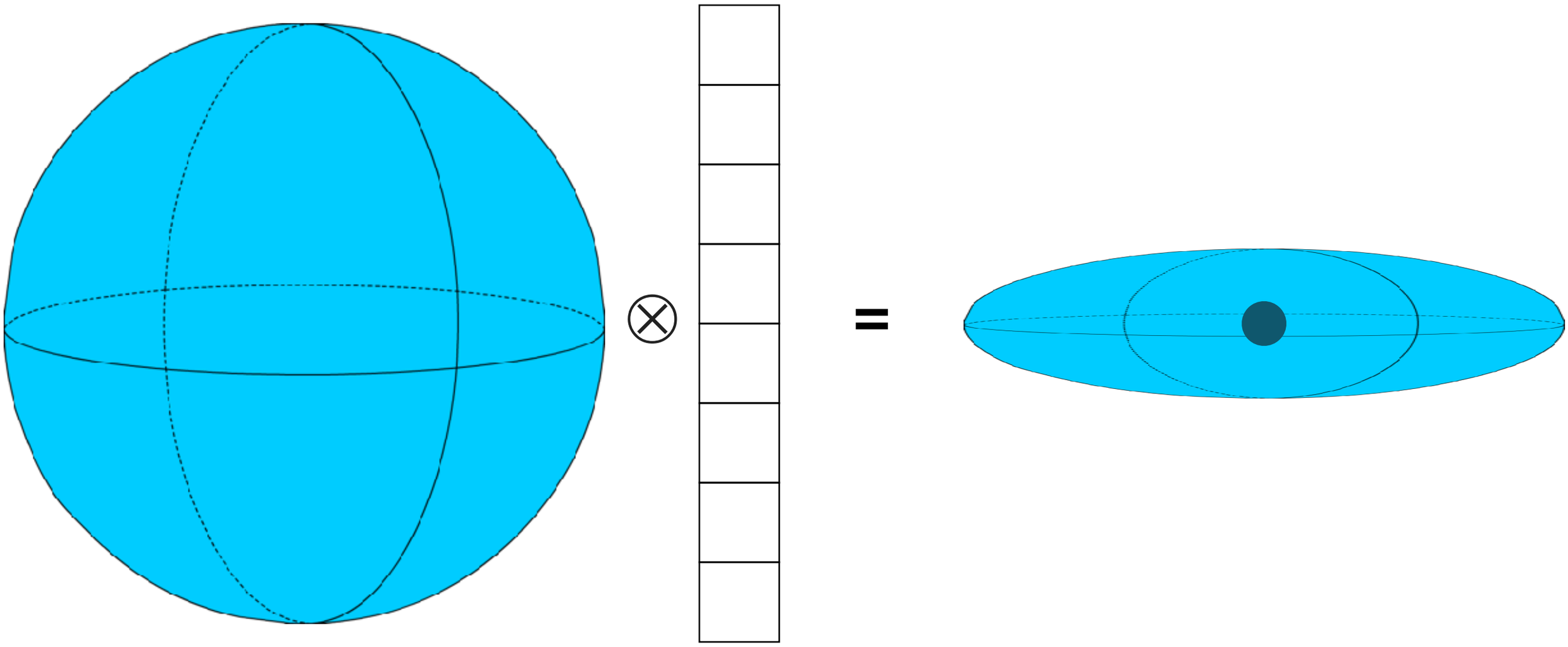
$$\mathcal{L}_{triplet} = \left[||f(a) - f(p)||_2^2 - ||f(a) - f(n)||_2^2 + m\right]_+$$



$$\mathcal{L}_{triplet} = \left[||f(a) - f(p)||_2^2 - ||f(a) - f(n)||_2^2 + m\right]_+$$

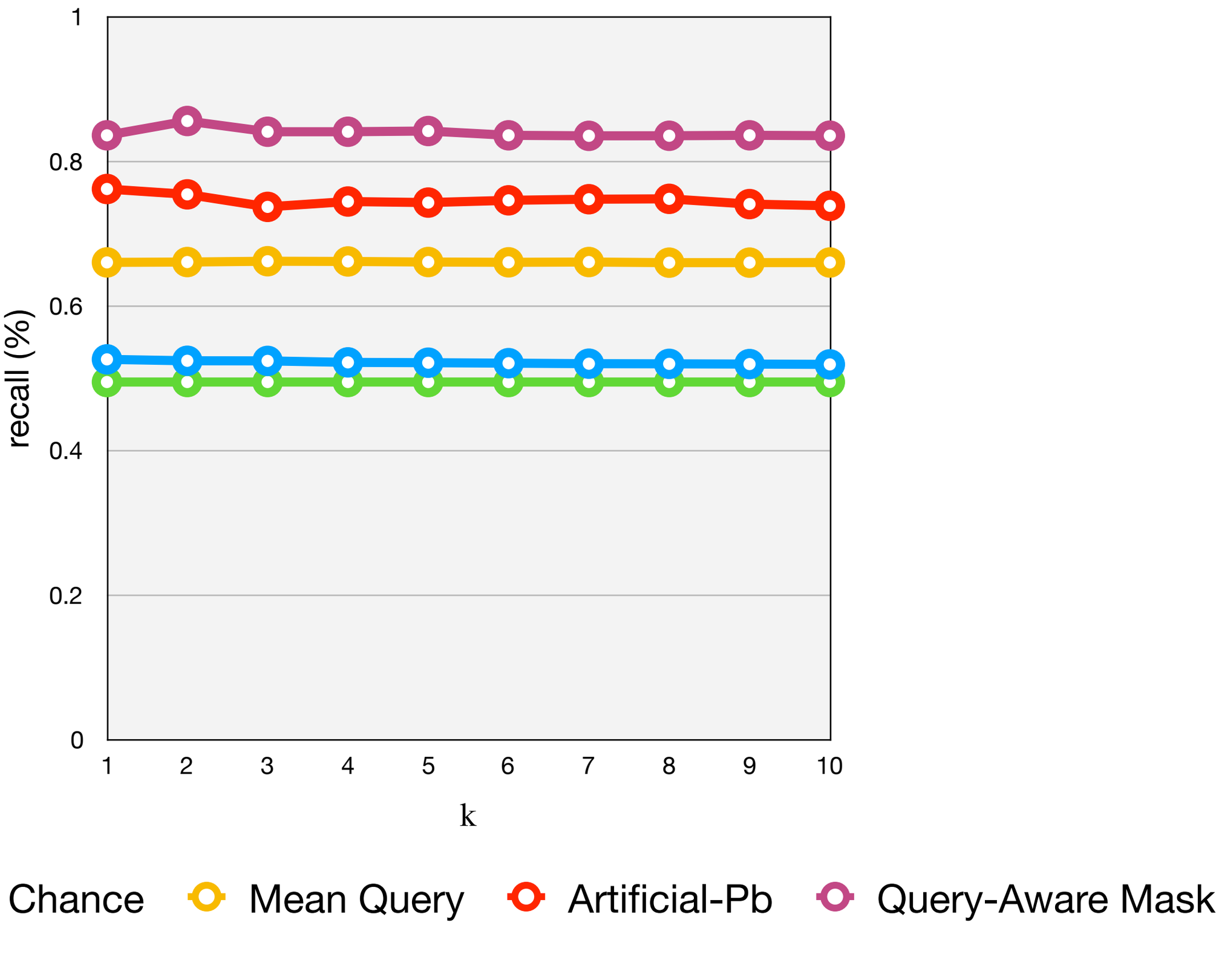


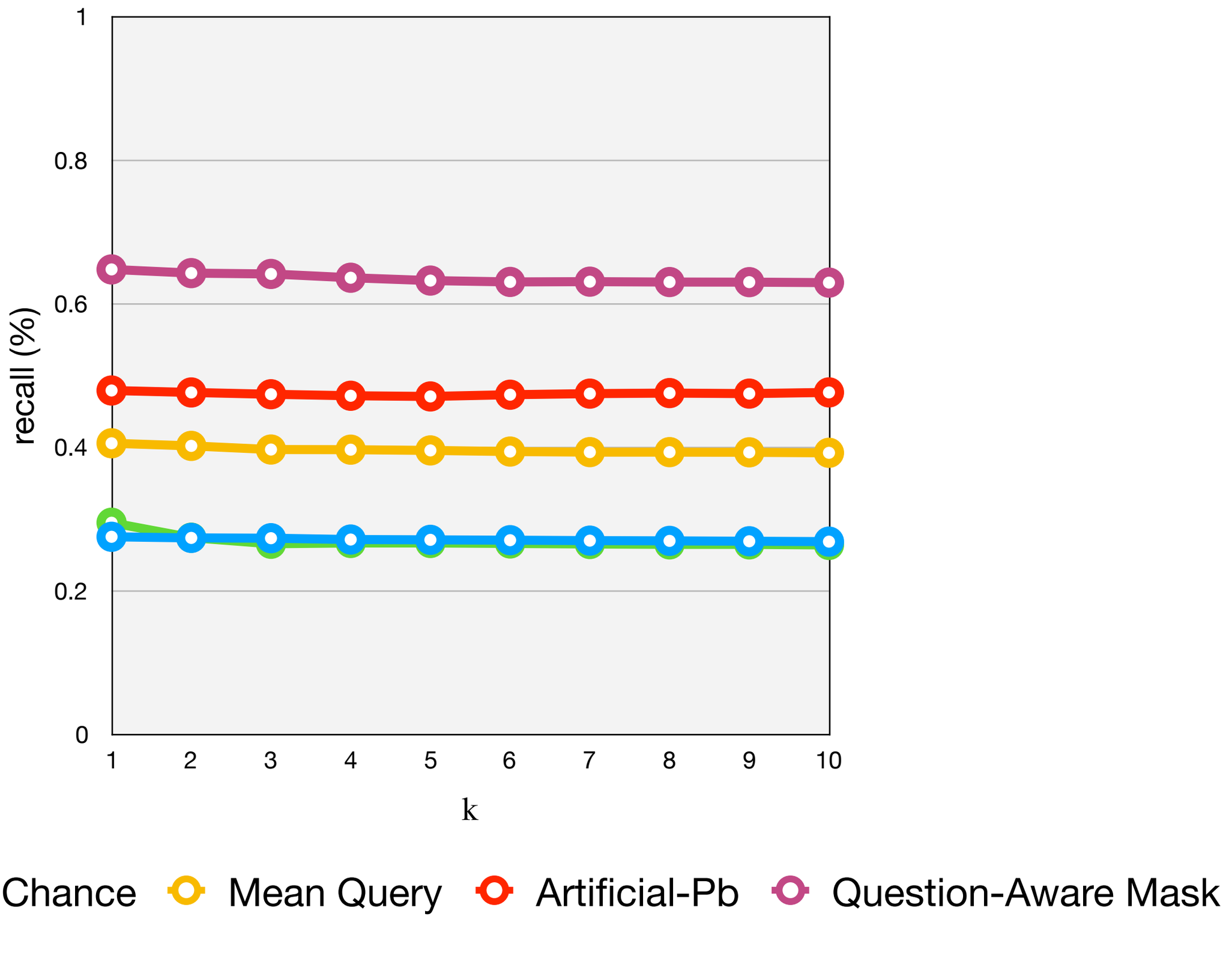
$$\mathcal{L}_{triplet} = \left[||f(a) - f(p)||_2^2 - ||f(a) - f(n)||_2^2 + m\right]_+$$



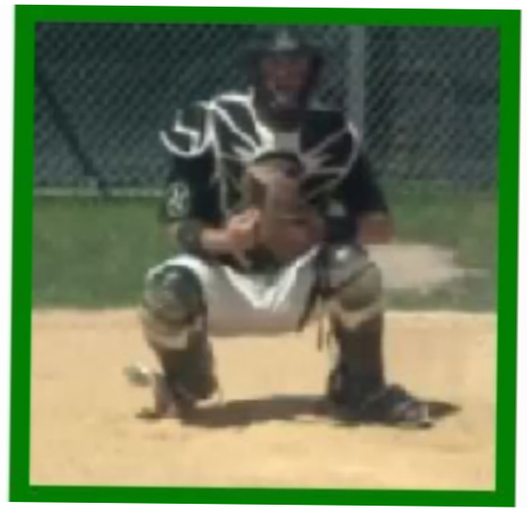
“right wrist **is above** head”

$$e'_{emb} = e_{emb} v_{mask}$$

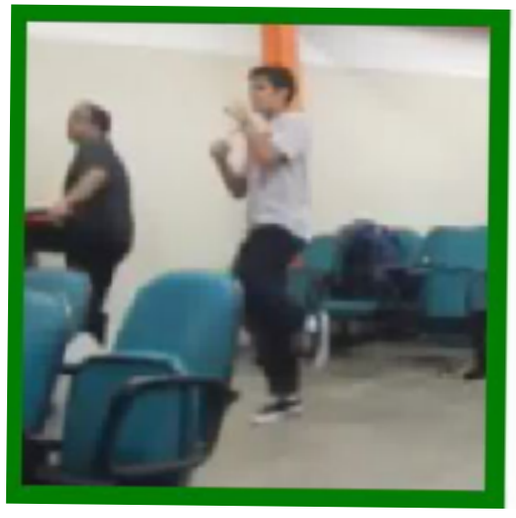

recall@k (1st order)
1
0.8
0.6
0.4
0.2
0
recall (%)
1
2
3
4
5
6
7
8
9
10
k
VGG-S
Chance
Mean Query
Artificial-Pb
Query-Aware Mask


recall@k (2nd order)
recall (%)
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
7
8
9
10
k
VGG-S
Chance
Mean Query
Artificial-Pb
Question-Aware Mask

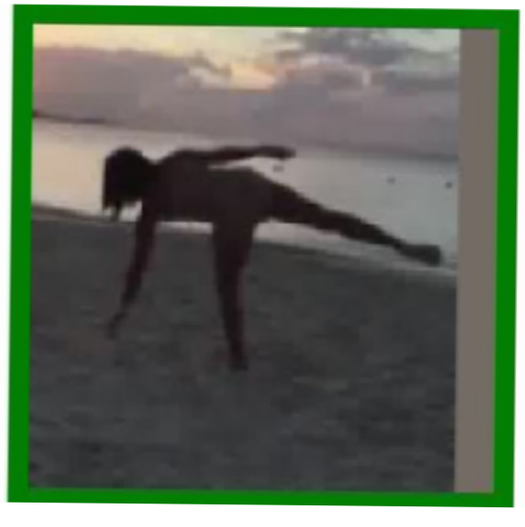
pelvis **is bent**



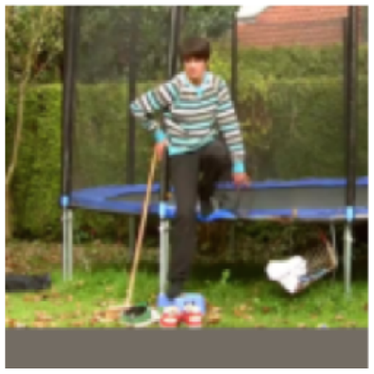
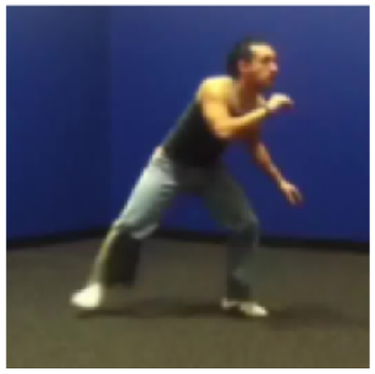
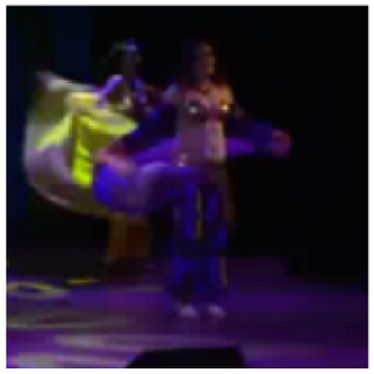
right wrist
is near
neck



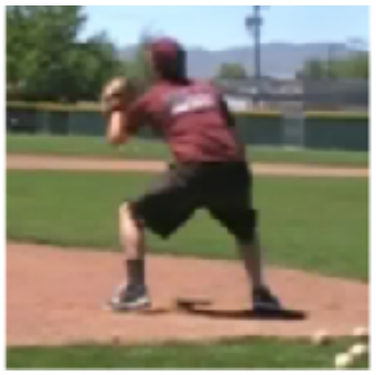
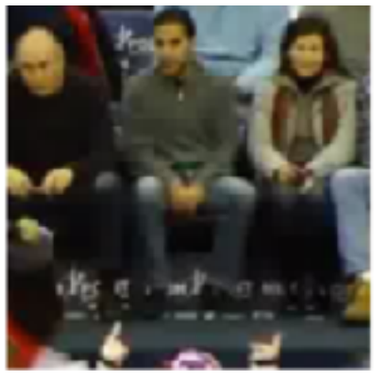
left wrist
is beyond
left elbow



pelvis **is bent**

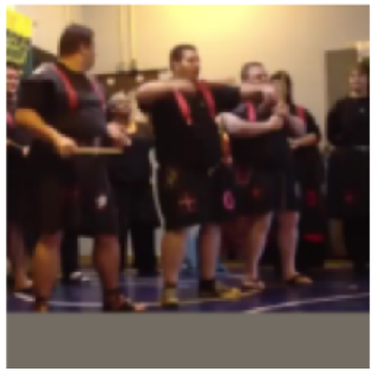





right knee
is far from
left knee

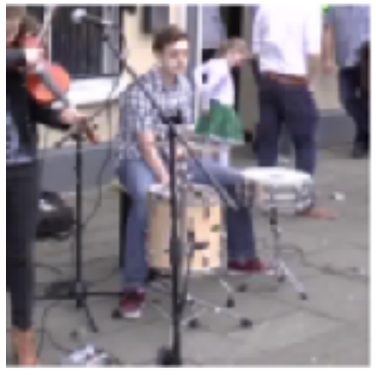
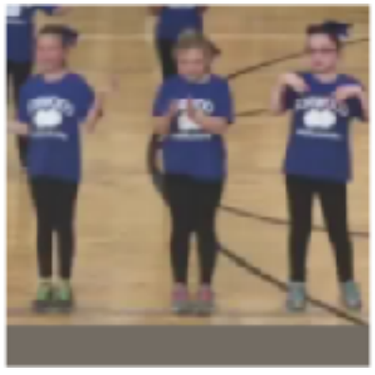




left wrist
is beyond
neck

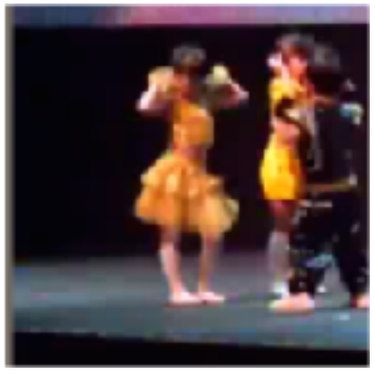


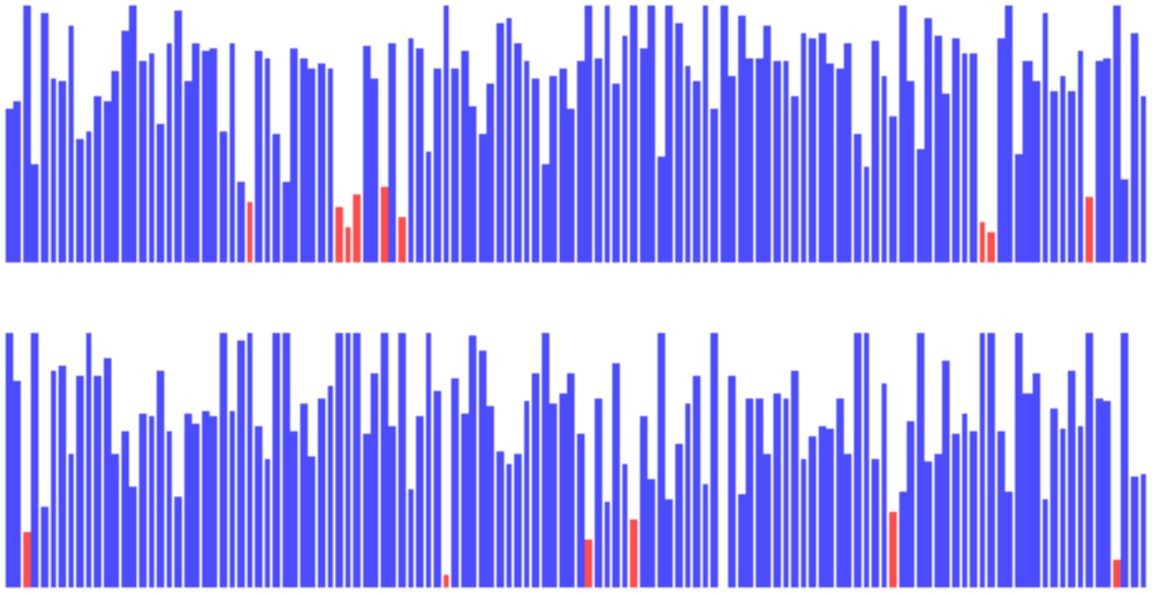

right elbow
is near
left elbow

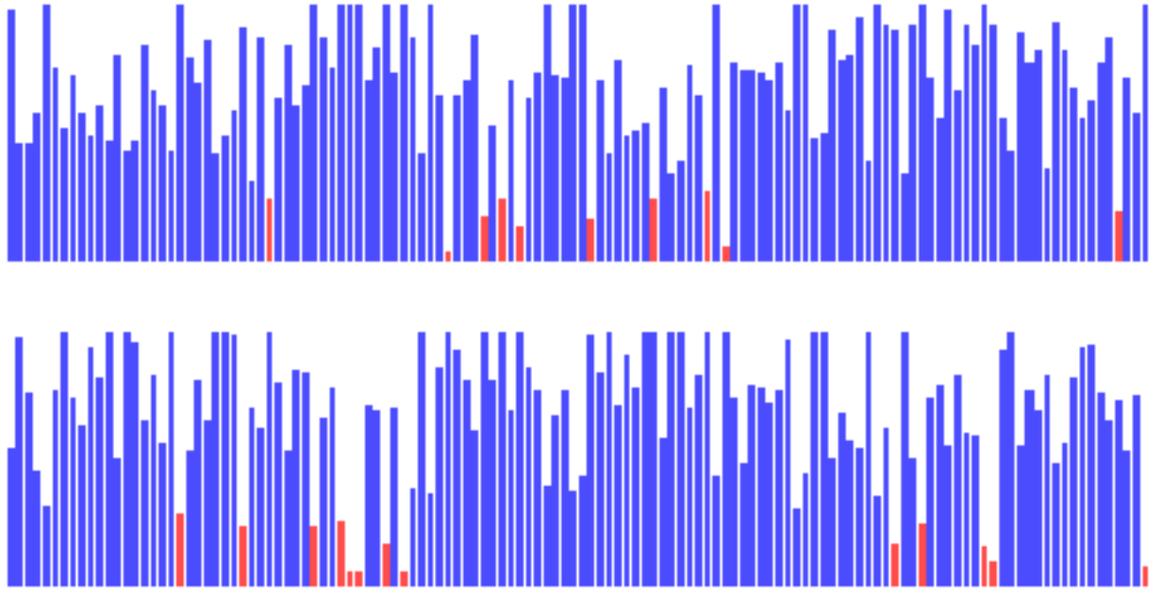




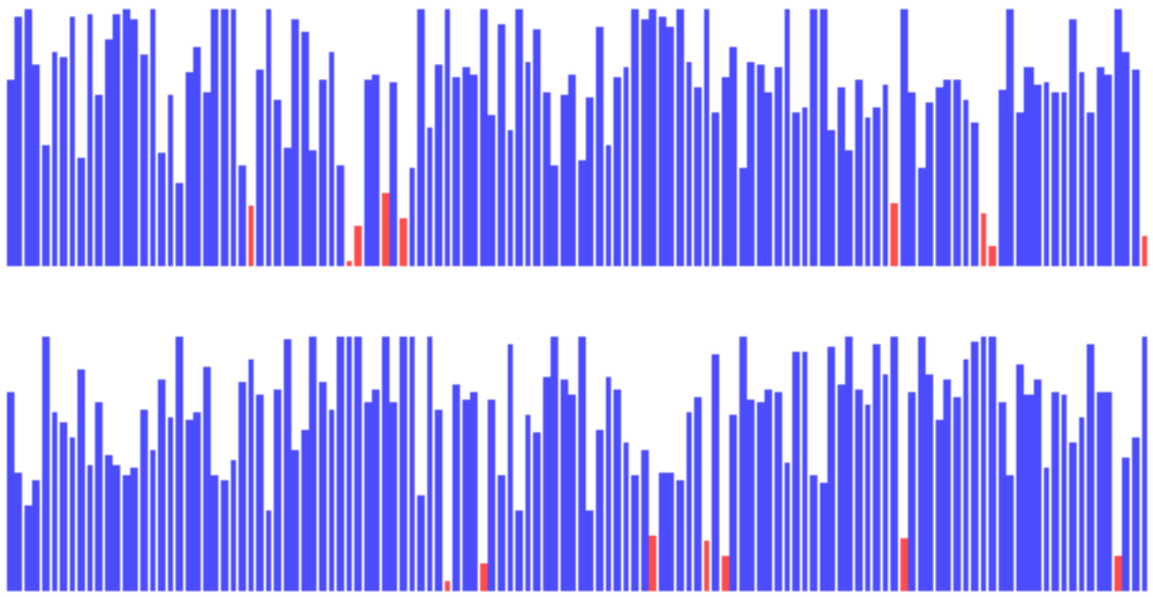
“right shoulder is (not) **bent**”



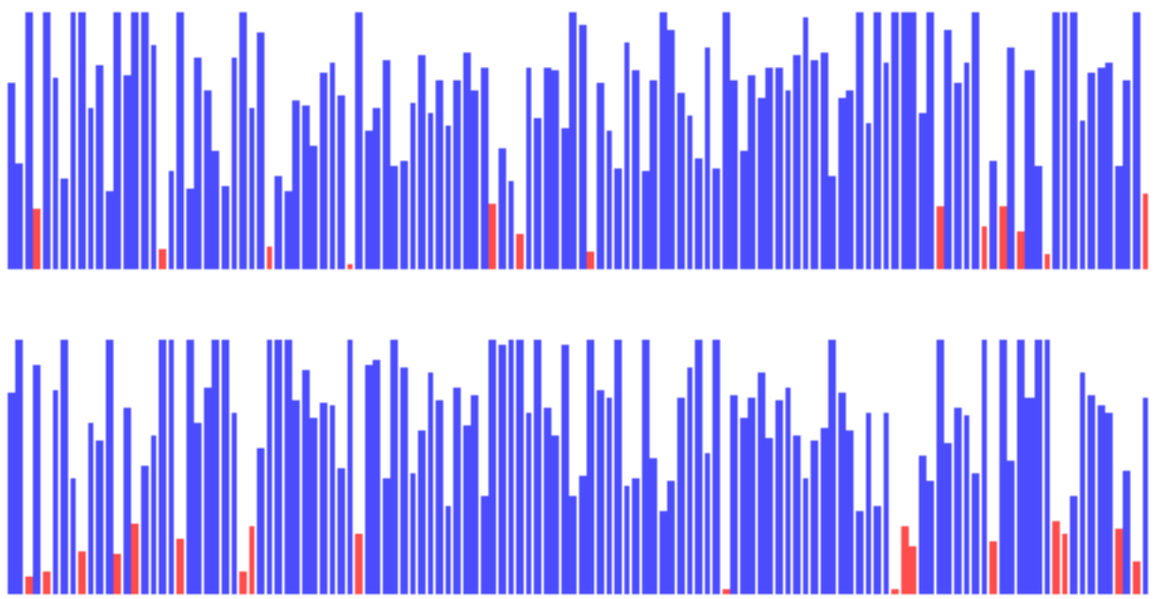
“left wrist is (not) **near** head”



“left wrist is (not) **beyond** pelvis”



“left knee is (not) **near** right knee”



Main Contributions

Pose-Aware embedding networks

Image-Language embedding network

Two solutions for language subsets

A dataset of 2D, 3D, and Language Primitive descriptors