Few-Shot Visual Relationship Co-Localization




Revant Teotia
IIT Jodhpur

Vaibhav Mishra
IIT Jodhpur

Mayank Maheshwari
IIT Jodhpur

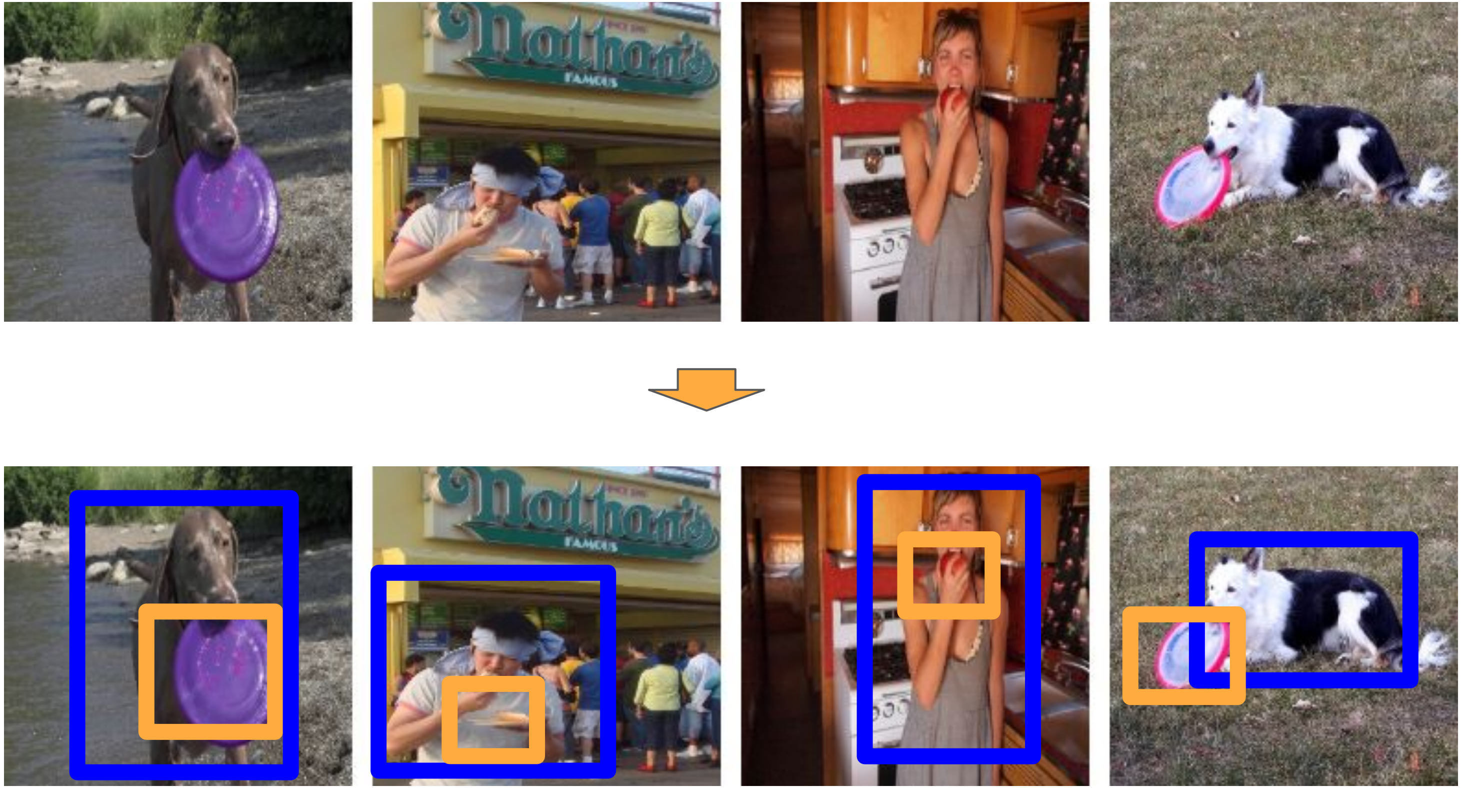
Anand Mishra
IIT Jodhpur

2021 ICCV OCTOBER 11-17 VIRTUAL

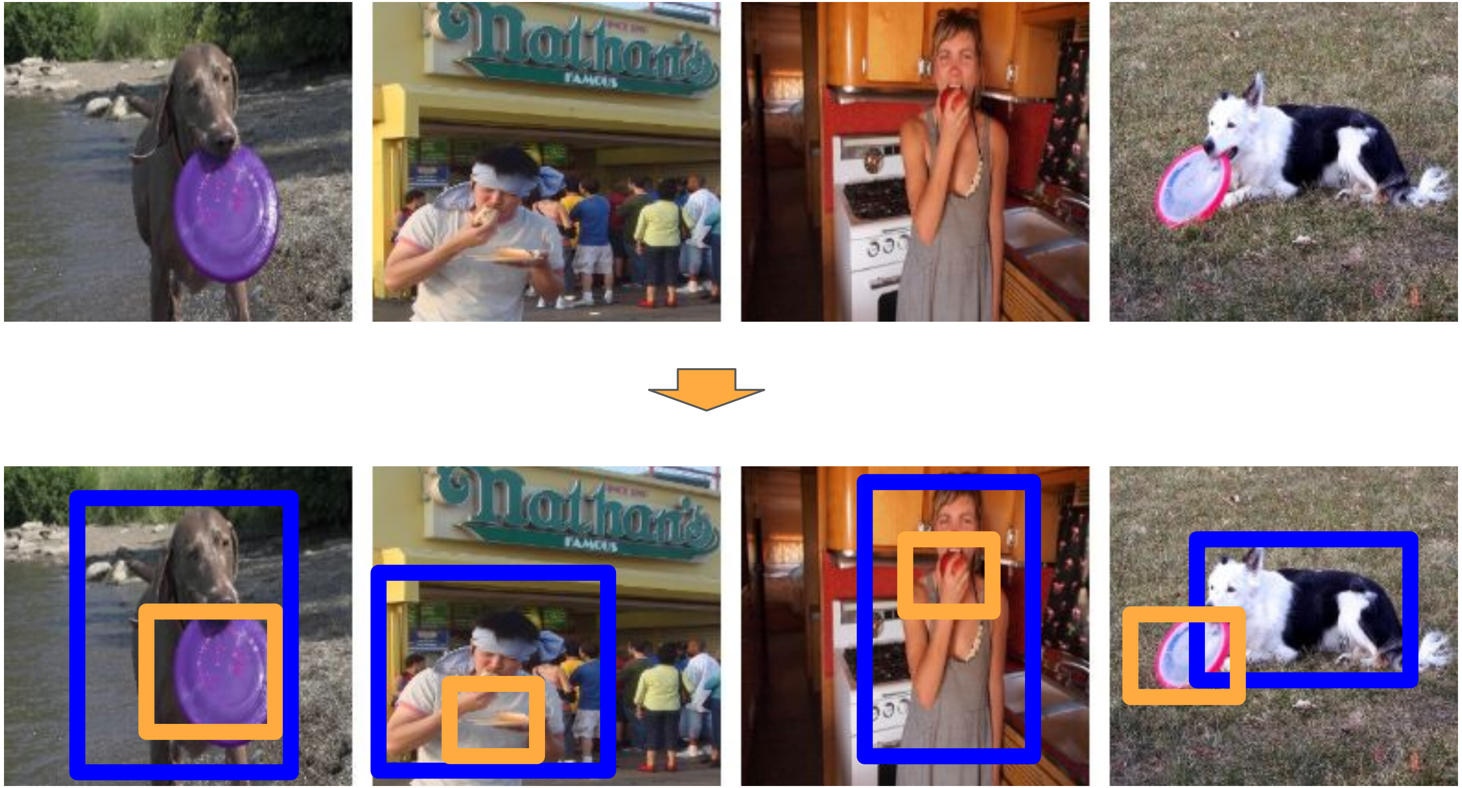
Indian Institute of Technology Jodhpur

॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥


Nathan's
FAMOUS


Nathan's
FAMOUS

Visual Relationship Co-localization



Visual Relationship = <Subject, Predicate, Object>

STEP 1

Forming VRC as a Labeling Problem

Forming VRC as a labelling problem → Getting the label set for an image → Computing pairwise cost → Episodic training → Inference Stage

Forming VRC as a labelling problem

Given a bag of images:

Image-2

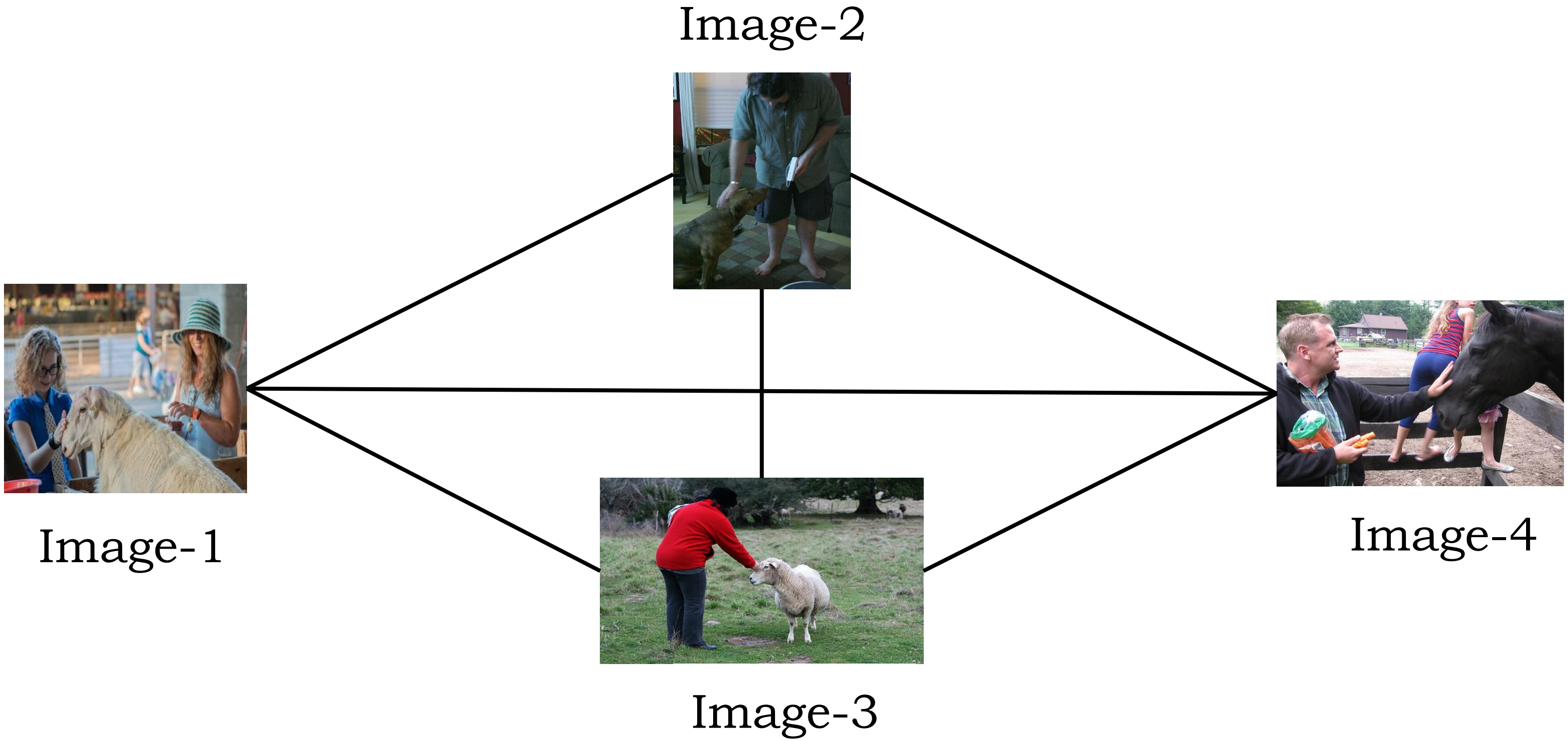
Image-1

Image-4

Image-3

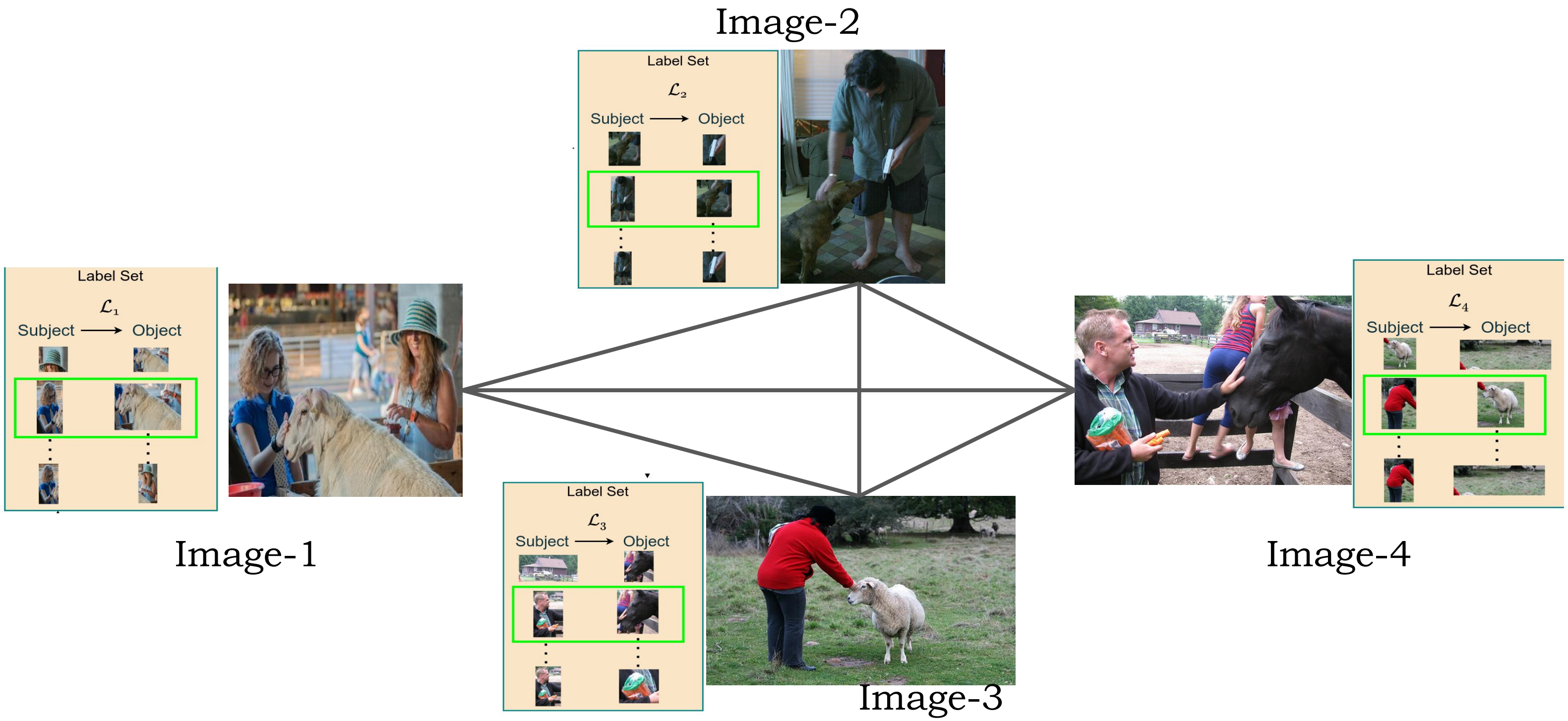
Forming VRC as a labelling problem

Construct a fully connected graph:



Forming VRC as a labelling problem

Label set = all possible visual relationships:

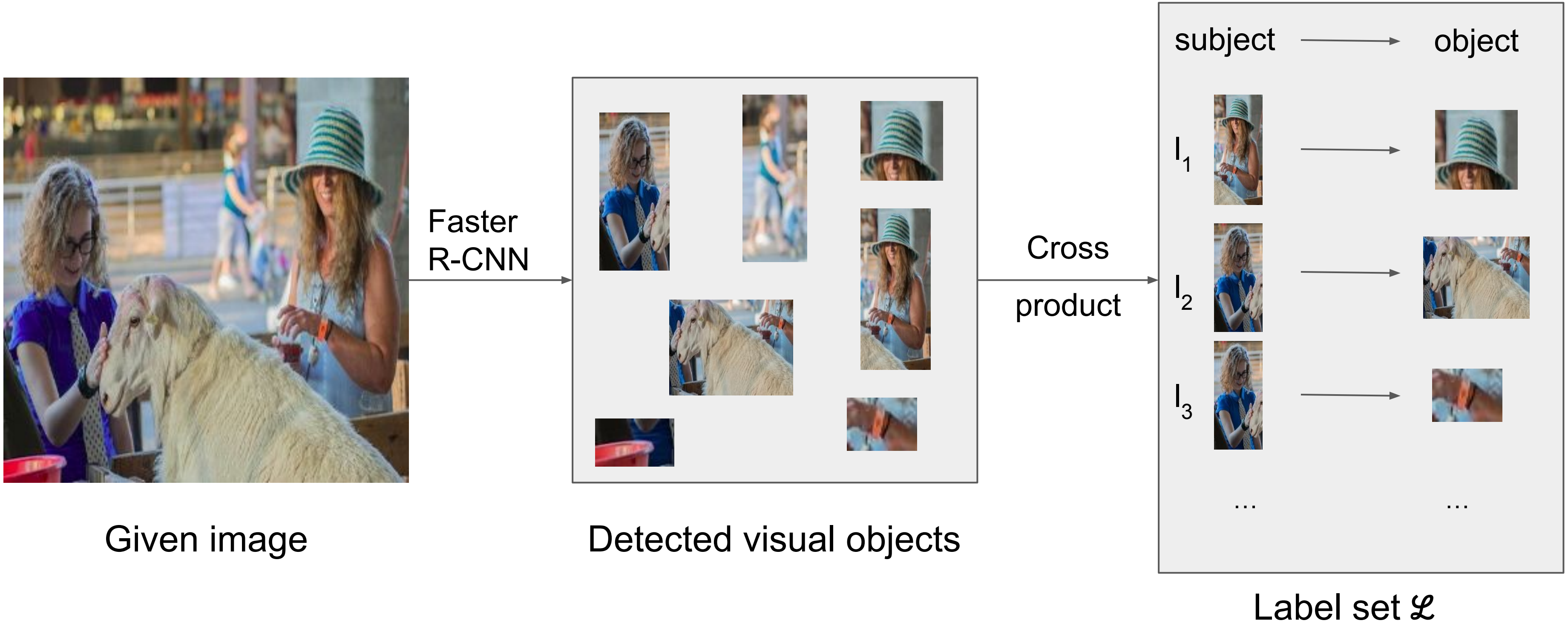


STEP 2

Getting the label set for an image

Forming VRC as a labelling problem → **Getting the label set for an image** → Computing pairwise cost → Episodic training → Inference Stage

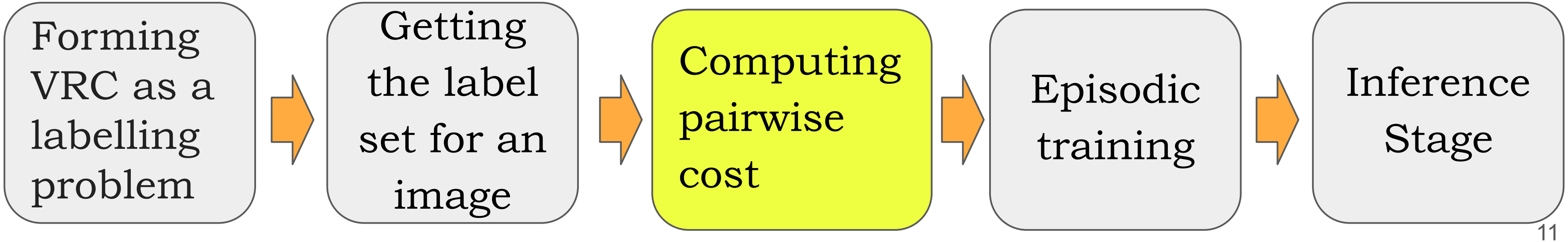
Getting the label set for an image



Label set = all possible visual relationships in an image
= all possible ordered pairs of detected visual objects in an image

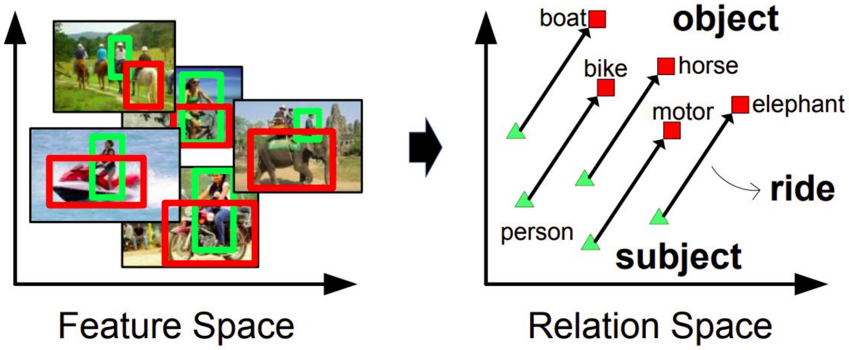
STEP 3

Computing pairwise cost



Computing pairwise cost

VTransE + Relation Network to learn VR similarity



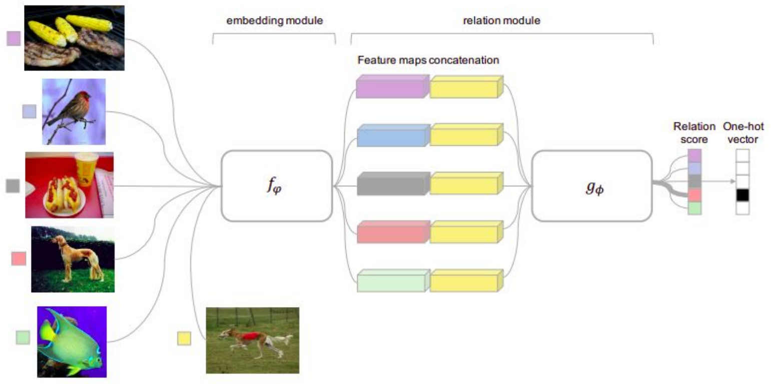


Figure 1: Relation Network architecture for a 5-way 1-shot problem with one query example.

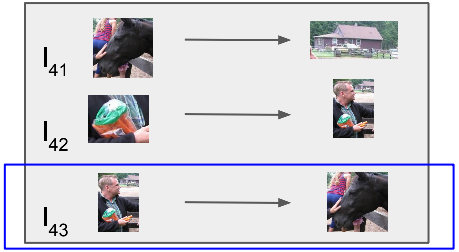
VTransE

[Zhang et al., CVPR 2017]

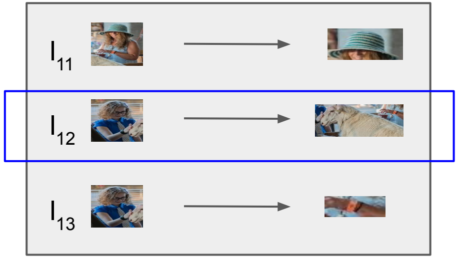
Relation Network

[Sung et al., CVPR 2018]

Computing pairwise cost

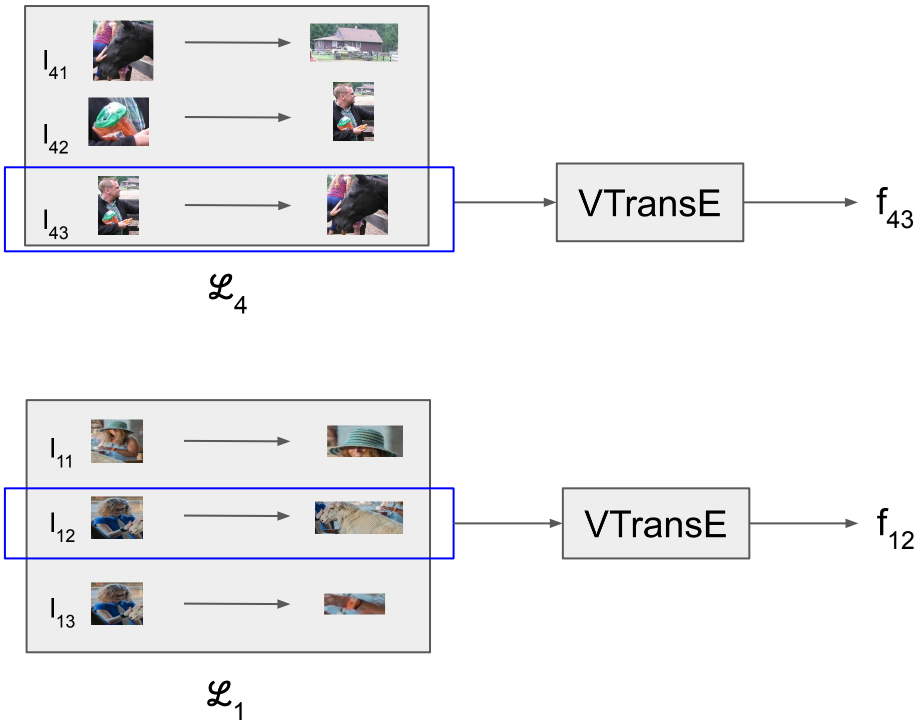


$\mathcal{L}_4$

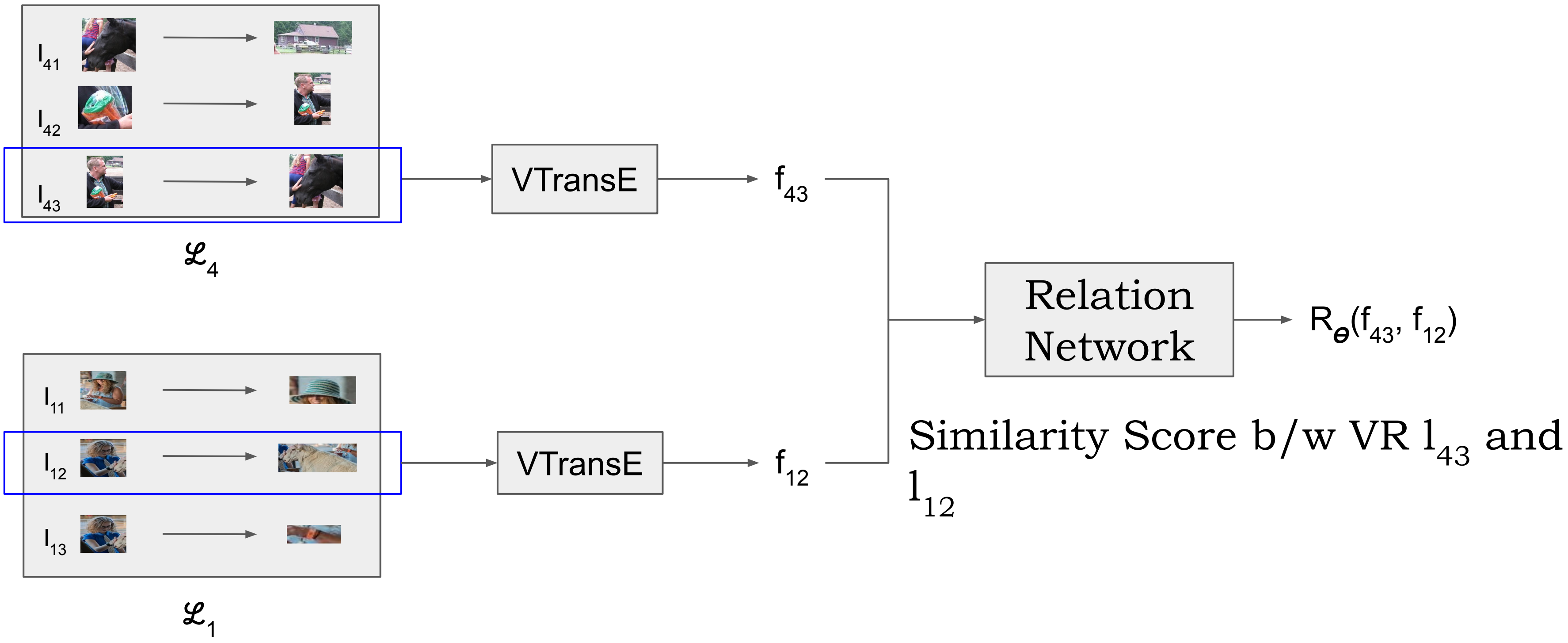


$\mathcal{L}_1$

Computing pairwise cost



Computing pairwise cost

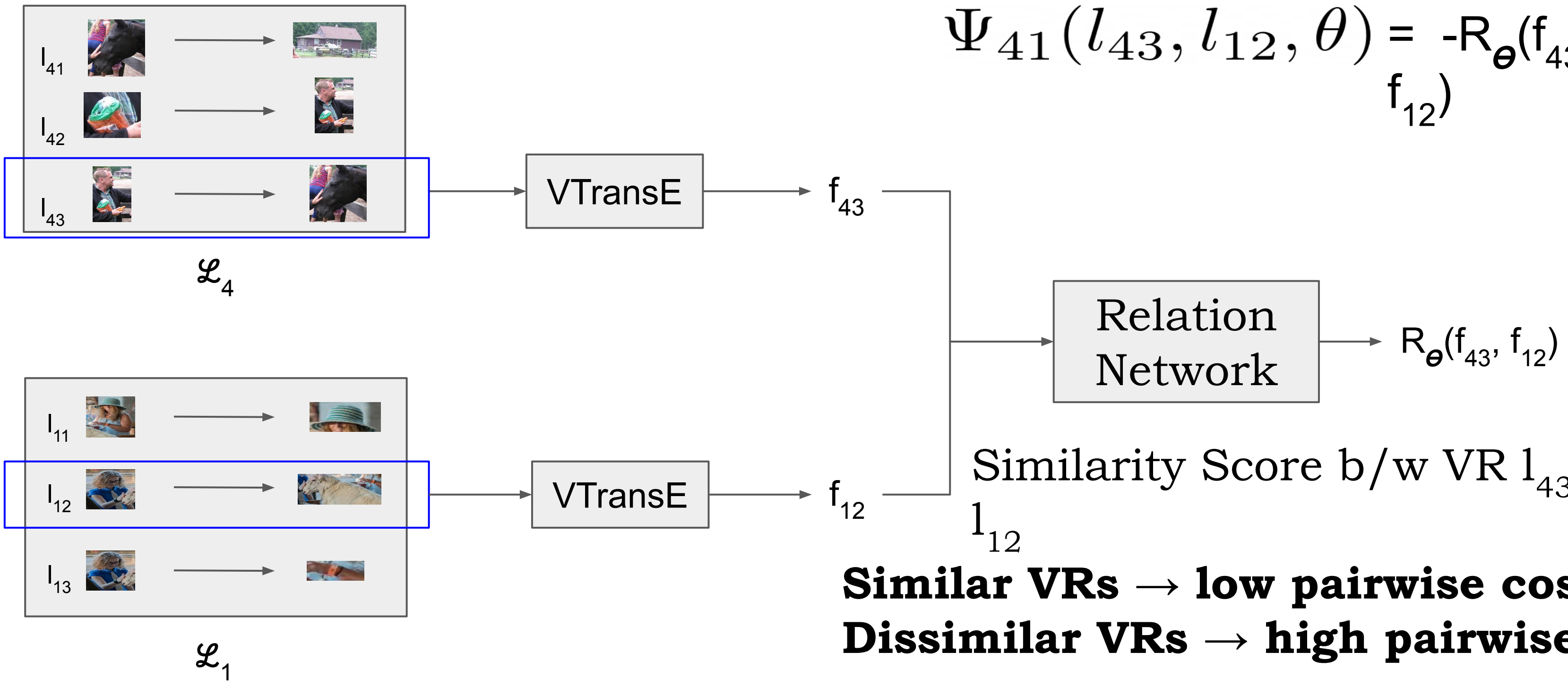


Computing pairwise cost

Pairwise Cost = - VR Similarity

Pairwise cost

$$\Psi_{41}(l_{43}, l_{12}, \theta) = -R_{\theta}(f_{43}, f_{12})$$

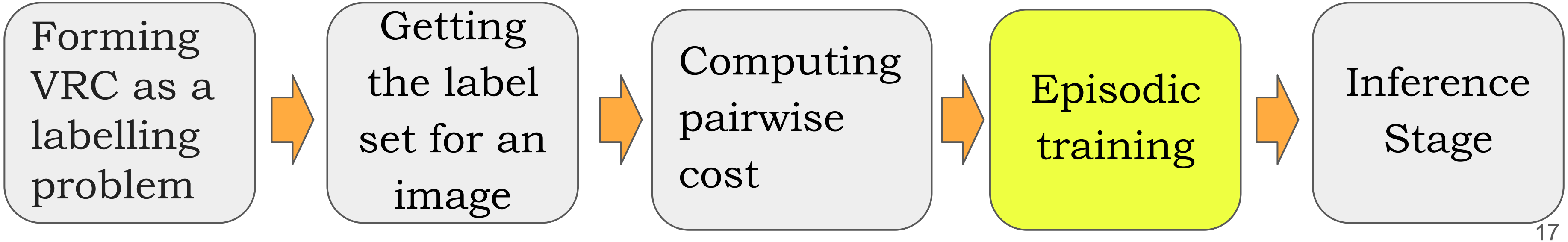


Similarity Score b/w VR l_{43} and l_{12}

Similar VRs → low pairwise cost
Dissimilar VRs → high pairwise cost

STEP 4

Episodic training



Episodic training

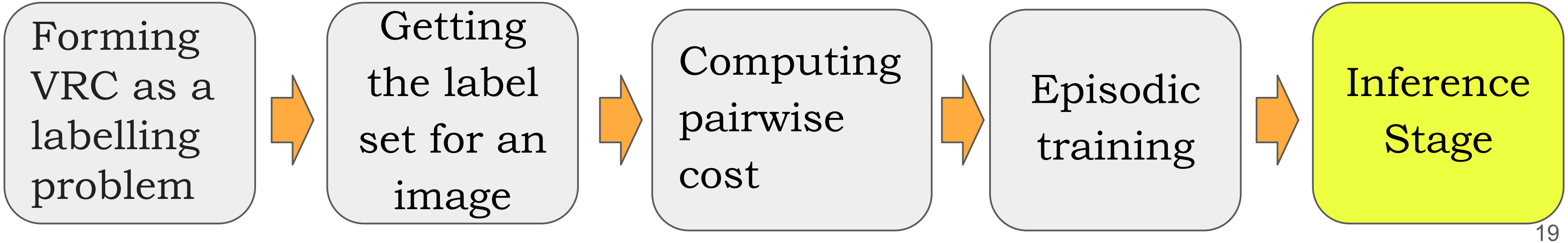
Episodic Training using Binary Log Regression Loss:

$$\text{Loss} = \frac{1}{N}\left(\sum_{(l_i,l_j)\in pos} L_p + \sum_{(l_i,l_j)\in neg} L_n\right)$$

where $L_p = \log\left(1 + e^{-R_\theta\left(f_{l_i}, f_{l_j}\right)}\right)$ and $L_n = \log\left(1 + e^{R_\theta\left(f_{l_i}, f_{l_j}\right)}\right)$

N : Total number of VR pairs created for a bag
pos : pairs with the common hidden predicate
neg : pairs with different predicate
R_Θ: Similarity computed using Relation Net

Step 5
Inference stage



Inference Stage

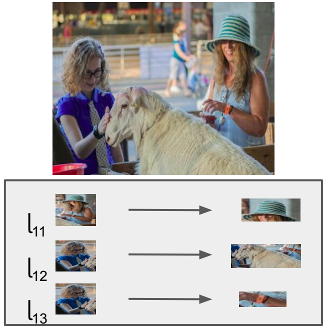


Image - 1

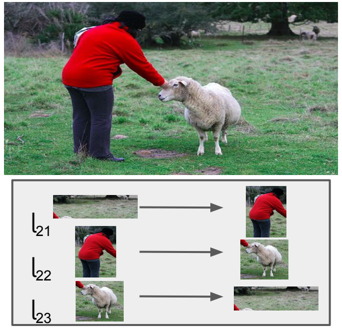


Image - 2

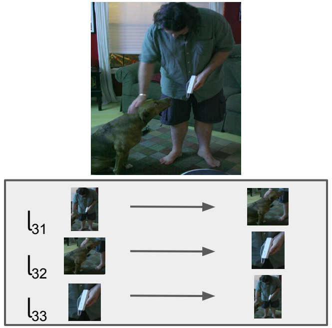


Image - 3

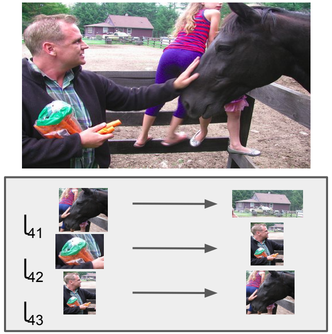
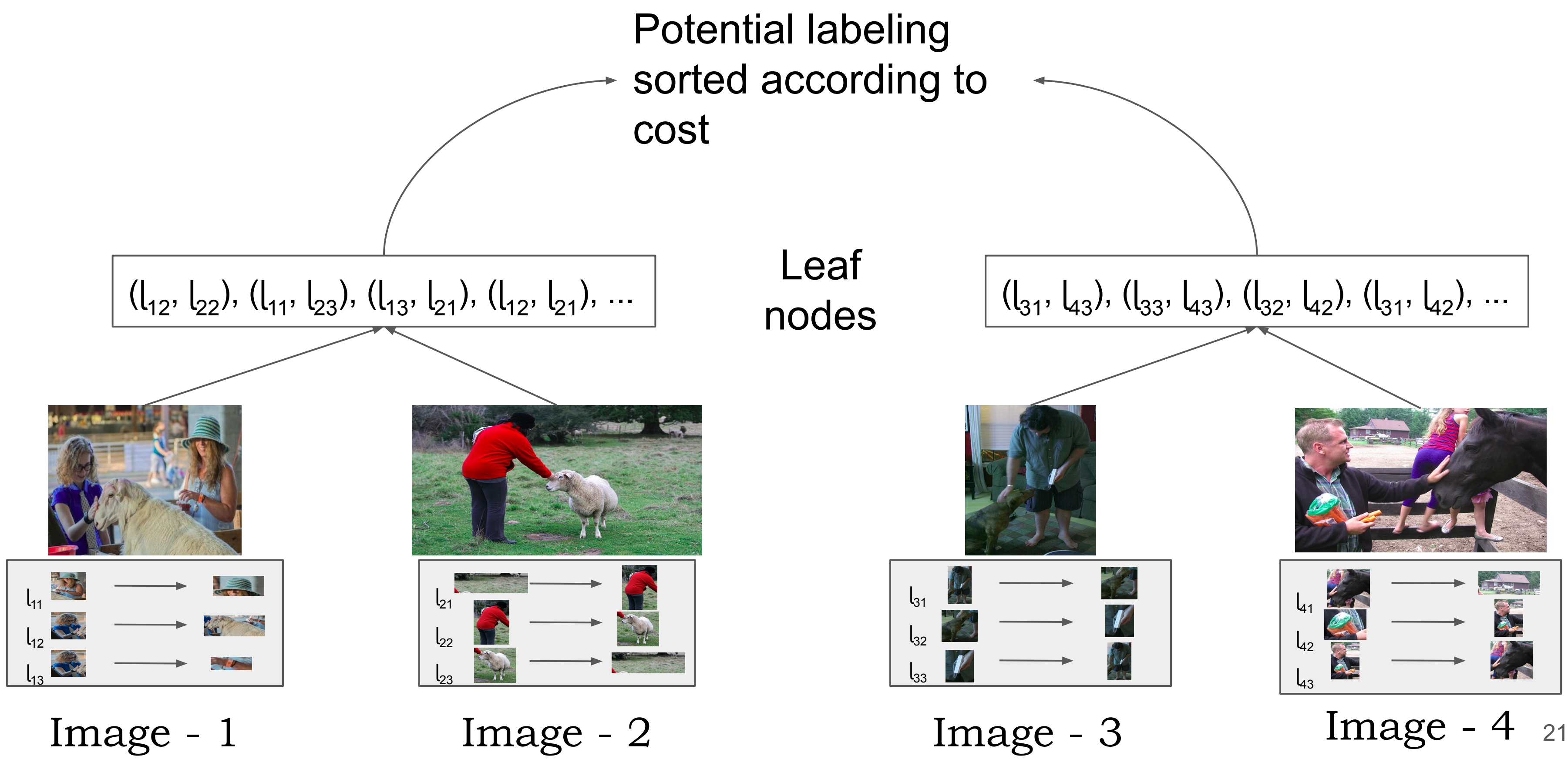
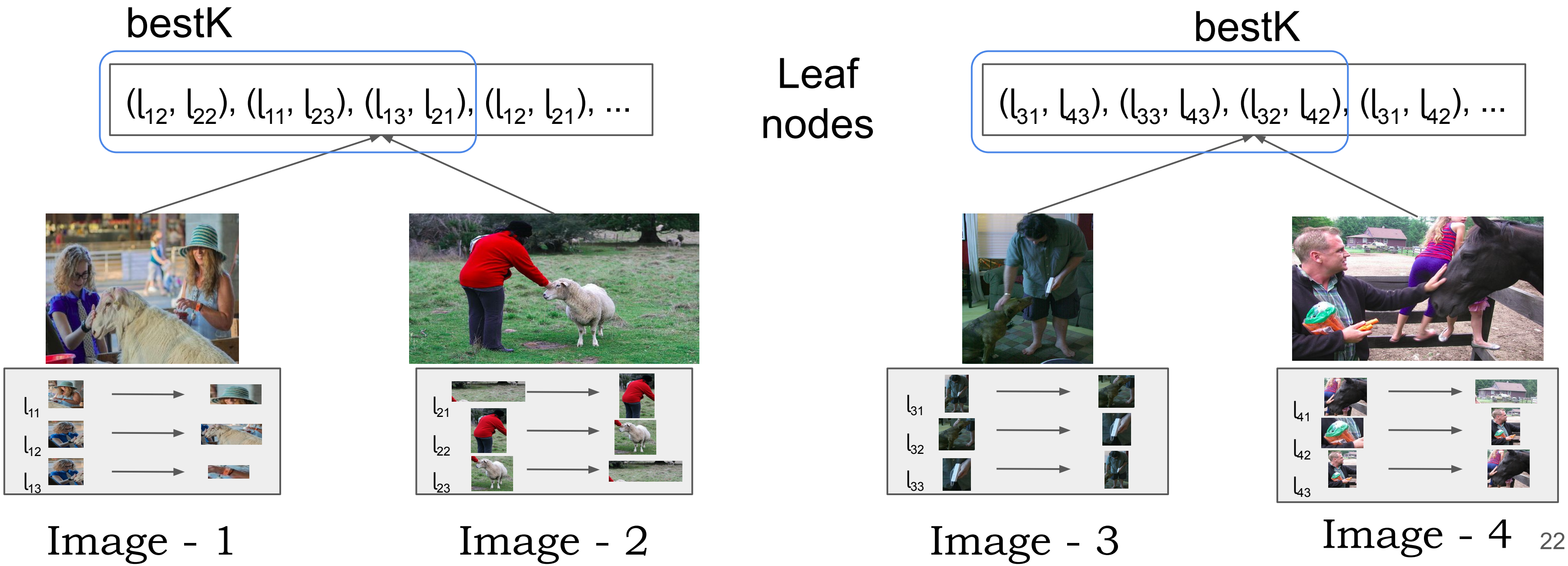


Image - 4

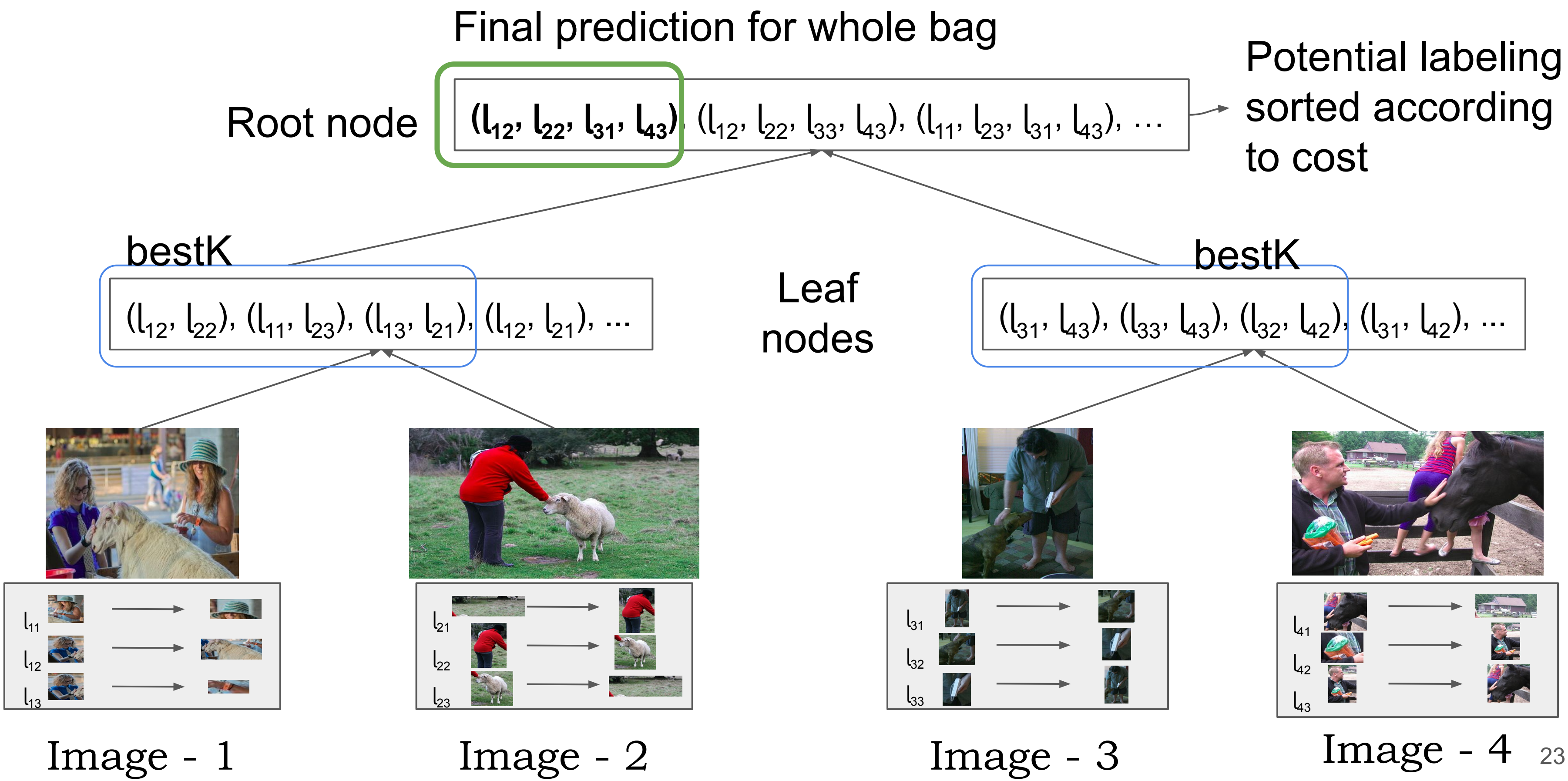
Inference Stage



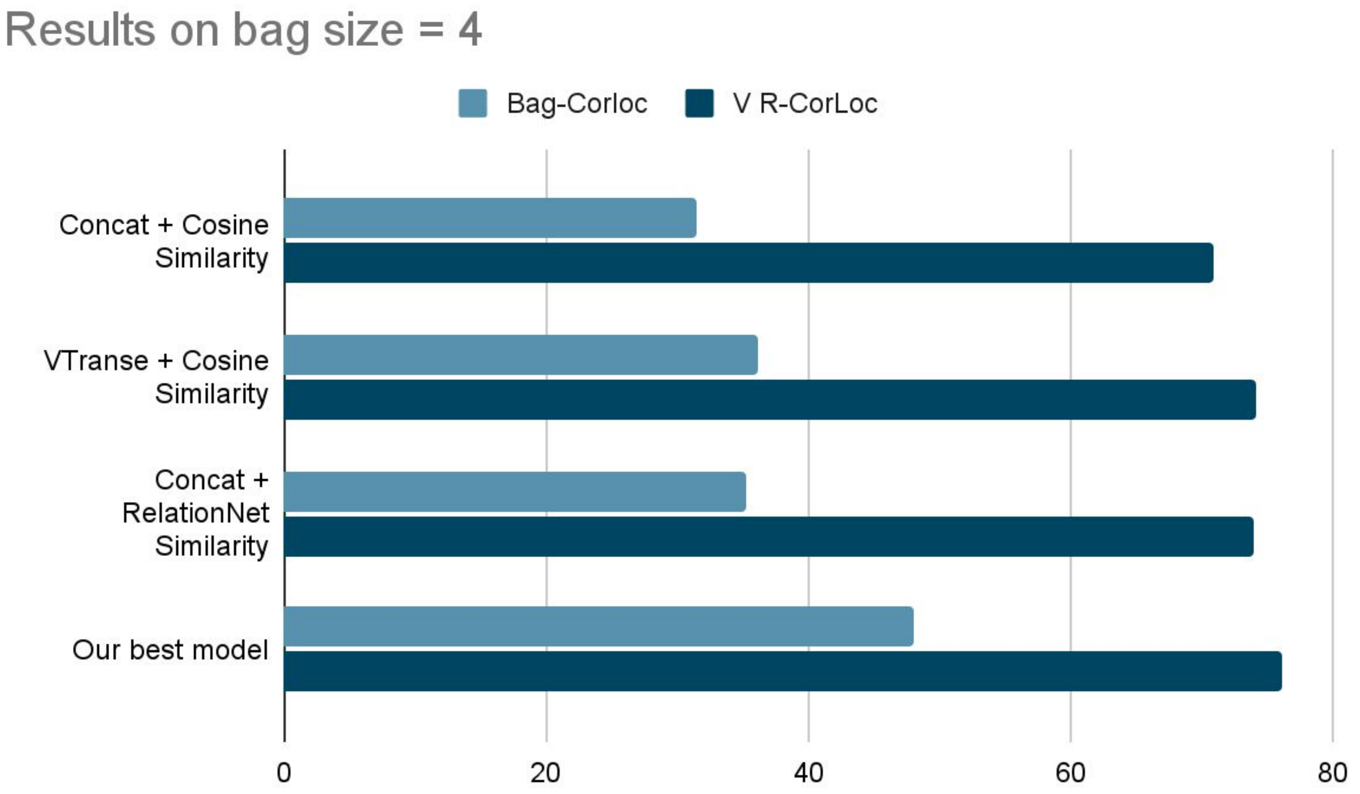
Inference Stage



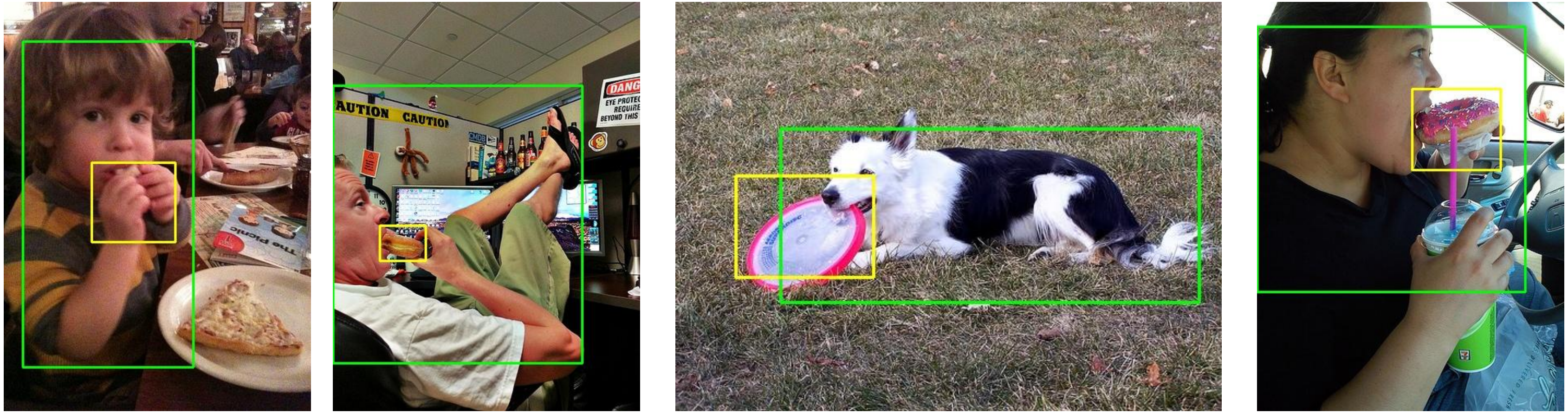
Inference Stage



Quantitative Results

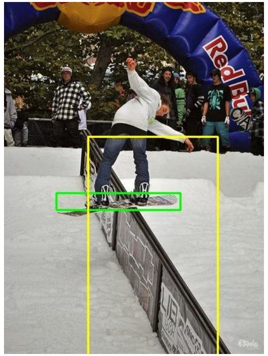


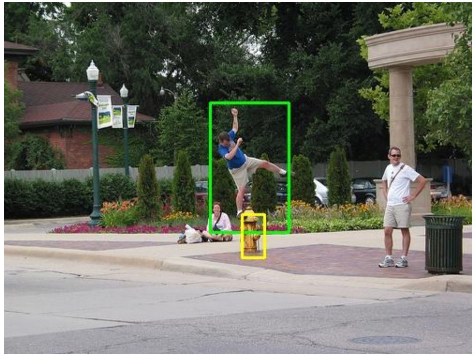
Qualitative Results



Latent visual relation: **Biting**

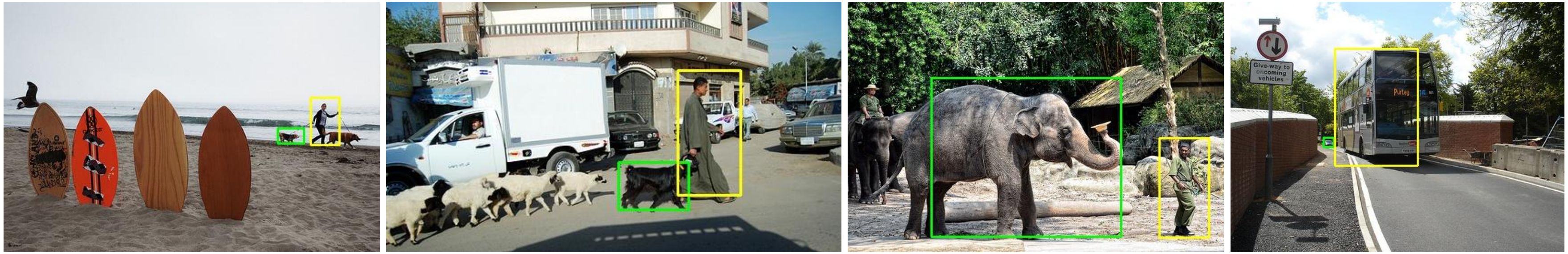
Qualitative Results



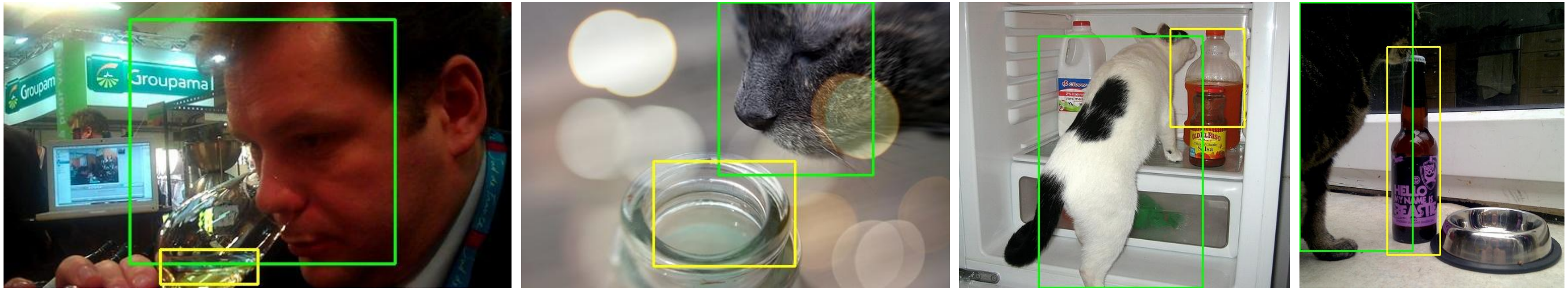
Latent visual relation: **Balancing On**

Qualitative Results



Latent visual relation: **Following**

Qualitative Results



Latent visual relation: **Sniffing**

Conclusion

- Visual Relationship Co-Localization: a novel task.
- A principled meta-learning based optimization framework
- Potential to open-up many future research avenues

Code Available!

Thank You

Getting the optimal labeling

$$\frac{1}{N}\left(\sum_{(l_i,l_j)\in pos} L_p + \right.$$

Episodic training with binary logistic regression loss :

For positive pairs: $L^p = \frac{1}{N_p}\sum_{(f_u,f_v)} (\log(1 + \exp(R_\Theta(f_u, f_v)))$

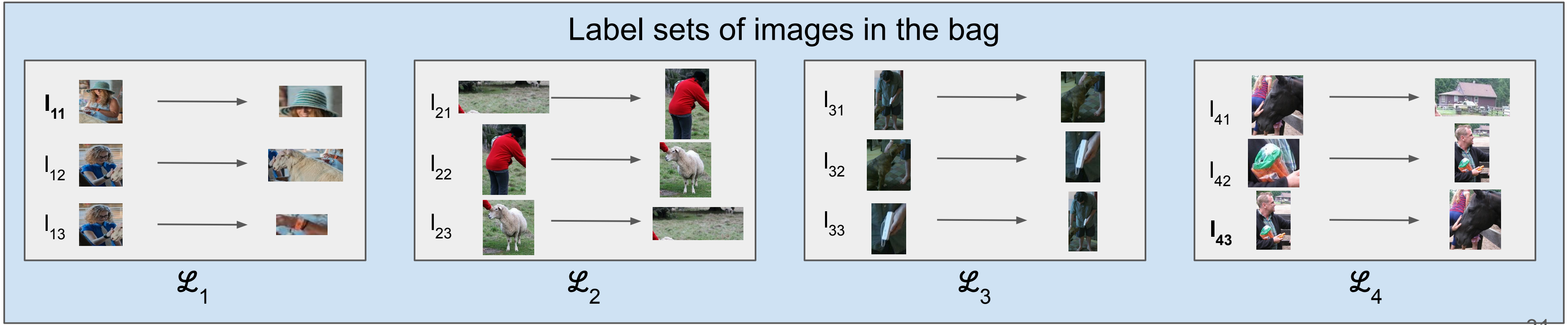
$$L_p = \log\left(1 + e^{-}\right.$$

Positive pairs = pair of labels / VRs sharing **common predicat**

For example : l_{22} = <woman, petting, sheep> and

l_{43} = <man, petting, horse>

$$L_n = \log\left(1 + e\right.$$



Getting the optimal labeling

Episodic training with binary logistic regression loss :

For negative pairs: $L^n = \frac{1}{N_n} \sum_{(f_u, f_v)} (\log(1 + \exp(R_\Theta(f_u, f_v))))$

Negative pairs = pair of labels / VRs having **different predicate**.

For example : l_{11} = <woman, wearing, hat> and

l_{43} = <man, petting, horse>

Label sets of images in the bag

Thank You

Any Questions?

Visual Relationship
= <Subject, Predicate, Object>

Visual Relationship
= <Subject, Predicate, Object>

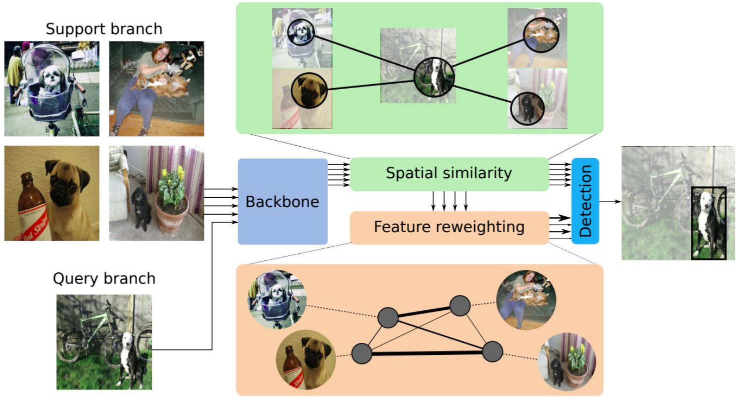
Visual Relationship
= <Subject, Predicate, Object>

Can you localize common visual relationships?

Visual Relationship Co-Localization : This work

Object Co-Localization and WSOL

[Shaban et al., ICCV 2019]



[Hu et al., ICCV 2019]

$$\Psi = \sum_{u=1}^{b} \left(\min_{t} \Psi_u(l_{ut}) + \sum_{v=1}^{b, u \neq v} \min_{t_1, t_2} \Psi_{uv}(l_{ut_1}, l_{vt_2}, \theta) \right)$$

$$\Psi = \boxed{\sum_{u=1}^{b}} \left(\min_{t} \Psi_u(l_{ut}) + \sum_{v=1}^{b, u \neq v} \min_{t_1, t_2} \Psi_{uv}(l_{ut_1}, l_{vt_2}, \theta) \right)$$

Sum over all the b images

$$\Psi = \sum_{u=1}^{b} \Big(\min_{t} \boxed{\Psi_u(l_{ut})} + \sum_{v=1}^{b, u \neq v} \min_{t_1, t_2} \Psi_{uv}(l_{ut_1}, l_{vt_2}, \theta) \Big)$$

Unary cost of assigning a label l_{ut} to image-u.
Considered uniform, does not contribute

$$\Psi = \sum_{u=1}^{b} \left(\min_{t} \Psi_u(l_{ut}) + \sum_{v=1}^{b, u \neq v} \min_{t_1, t_2} \boxed{\Psi_{uv}(l_{ut_1}, l_{vt_2}, \theta)} \right)$$

Pairwise cost of assigning labels l_{ut1} to image u and l_{vt2} to image v.

Lower when predicates of l_{ut1} and l_{vt2} are **semantically similar**

Todo list and extra slides next

TODO list :

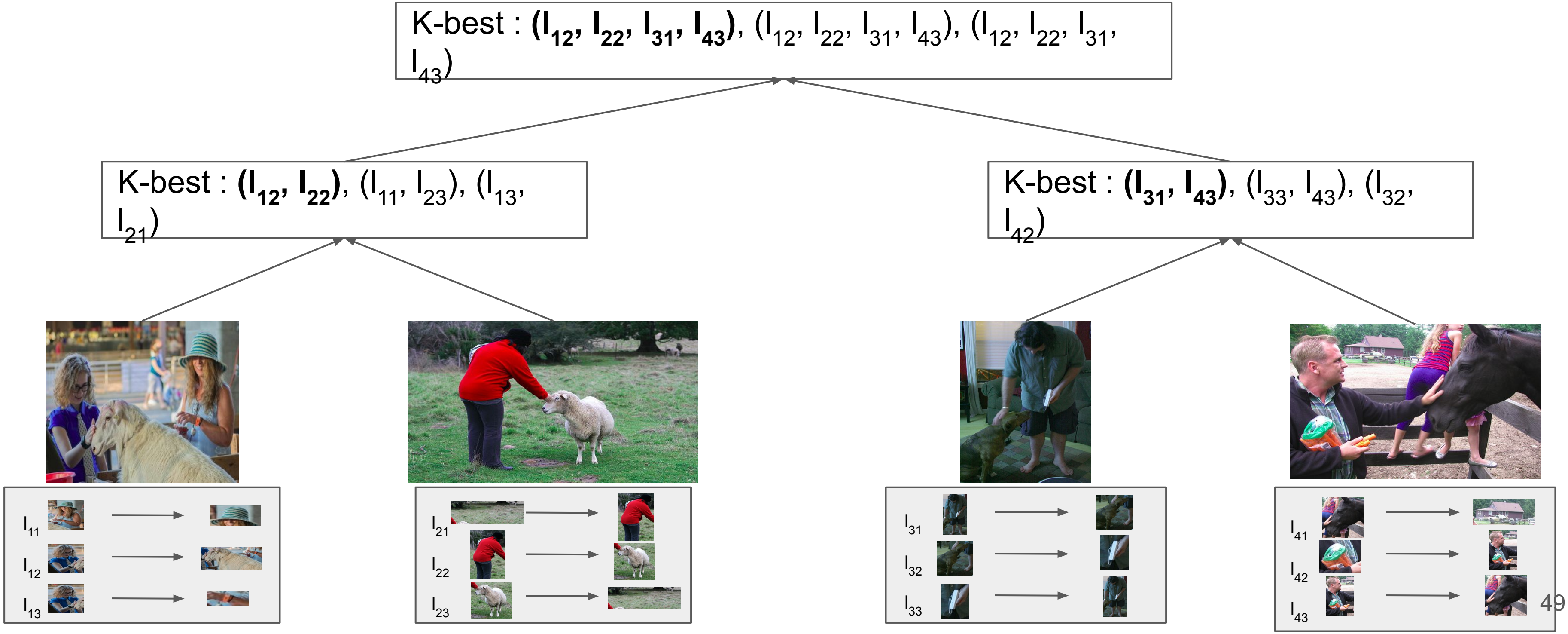
1. ~~First 2 slides : 1 problem statement + related work + why is a few-shot way … : [Mayank]~~
2. ~~3rd + 4th slide : How graph labeling using potential function~~
3. ~~2 slides probably : How are labels sets created(almost done) + RelationNet to find similarity/cost (almost done)~~
4. ~~Show how to train RelationNet in an episodic way : done~~ : NEEDS MORE WORK
5. ~~Inference algorithm : (almost done) finetune figure + how to explain :~~
6. Performance metrics : Mayank (can explain better)
7. Results (quantitative + visual results) : Mayank and Vaibhav

Speaker notes : use those

Keeping latex equation just in case

L^p = \frac 1 N_p \Sigma_{(f_u, f_v)} (log(1+ exp(- R_\Theta (f_u, f_v))))

Inference algorithm : dividing large bag into smaller ones
--> solve smaller subproblems : combine smaller solutions

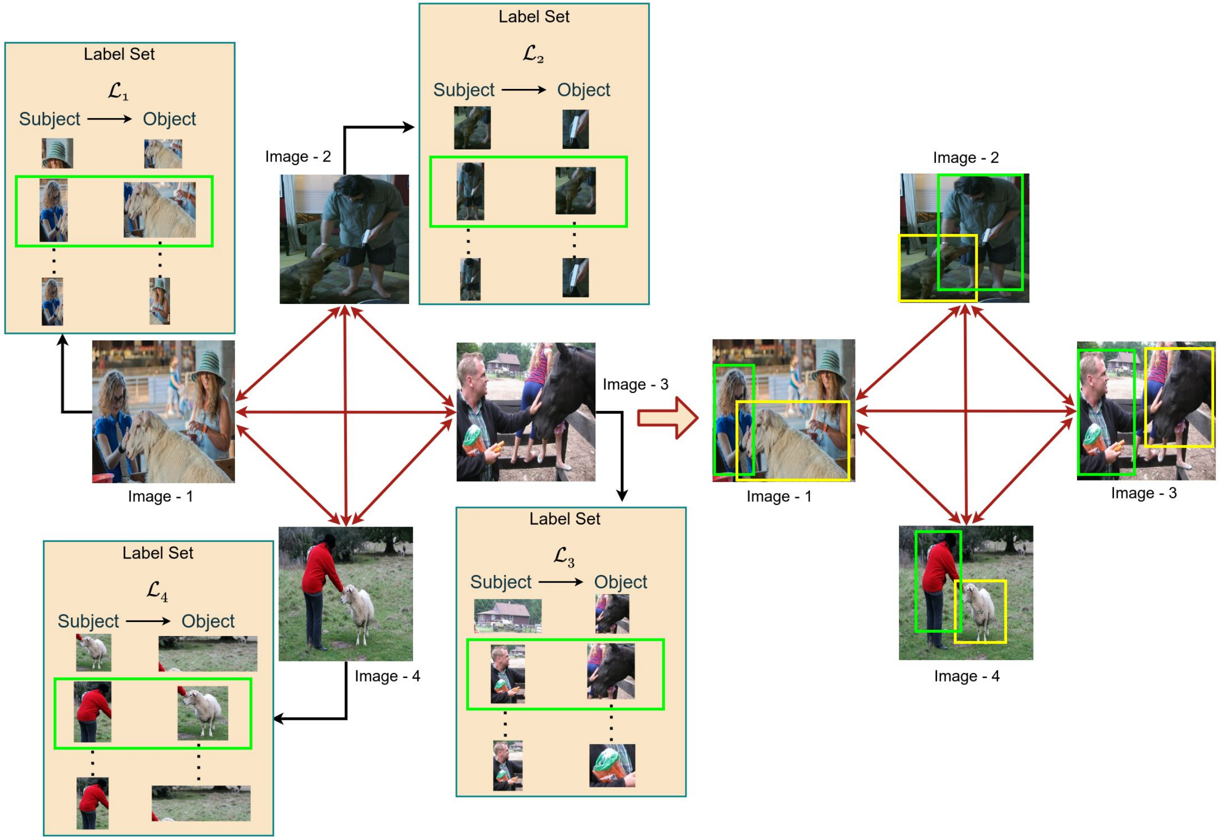


Problem Formulation : how a graph labelling problem : **what is the cost function**

Cost function

$$\Psi = \sum_{u=1}^{b} \left(\min_{t} \Psi_u(l_{ut}) + \sum_{v=1}^{b, u \neq v} \min_{t_1, t_2} \Psi_{uv}(l_{ut_1}, l_{vt_2}, \theta) \right)$$

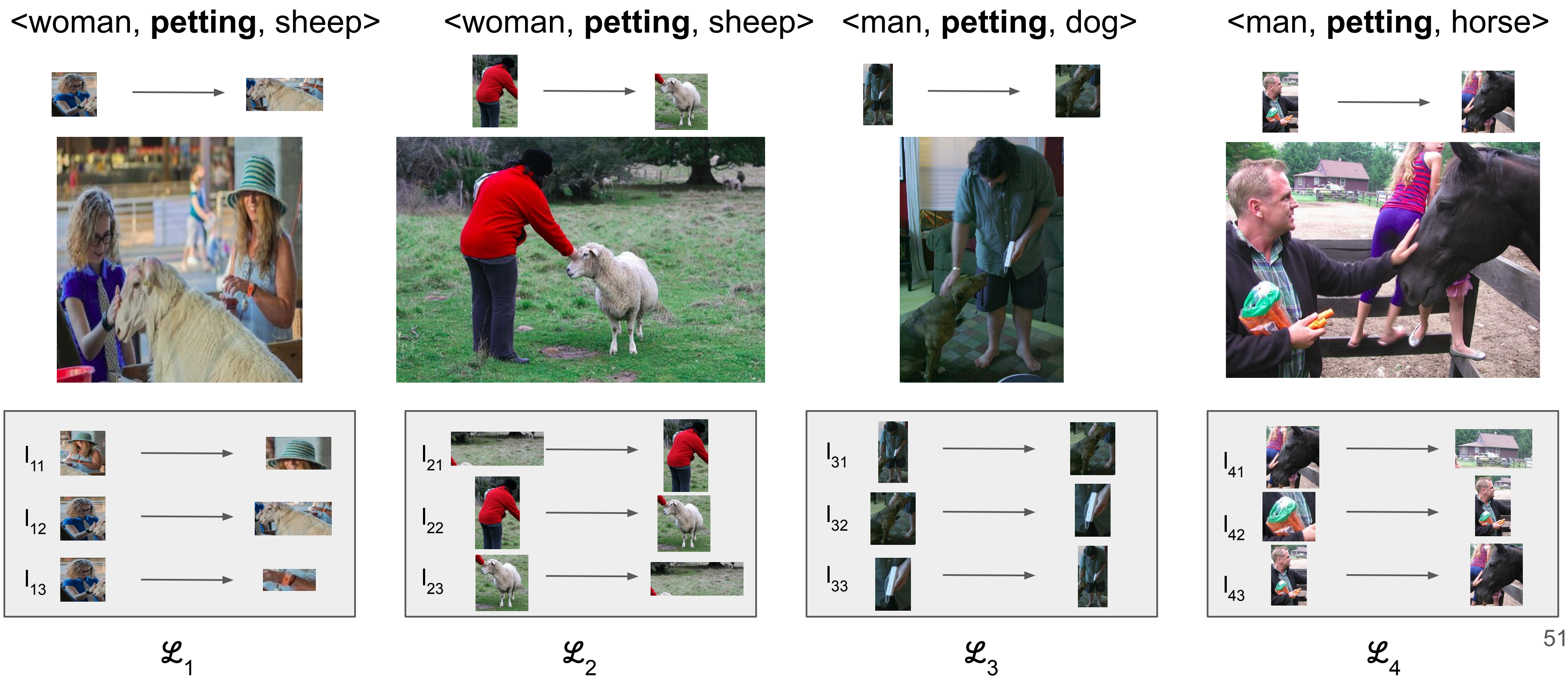
Unary cost : $\Psi_u(l_{ut})$



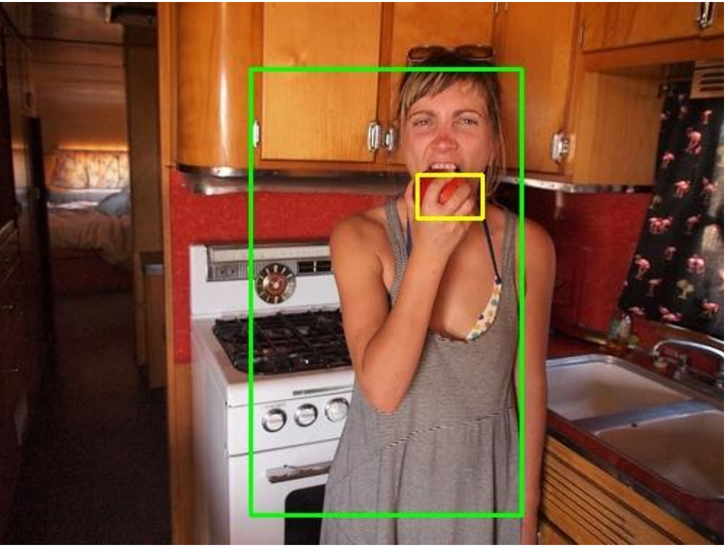
Bag of 4 images with common latent predicate = “petting”

Optimal selection (O) = (l_{1x}, l_{2x}, l_{3x}, l_{4x}), where $l_{ix} \in \mathcal{L}_i$ s.t. And all selected labels / visual relationships have **same predicate**.

For this illustration : O = (l_{12}, l_{22}, l_{31}, l_{43}) and the common hidden predicate = “petting”



How are we localizing a VR in this work. : [need to show this but where]



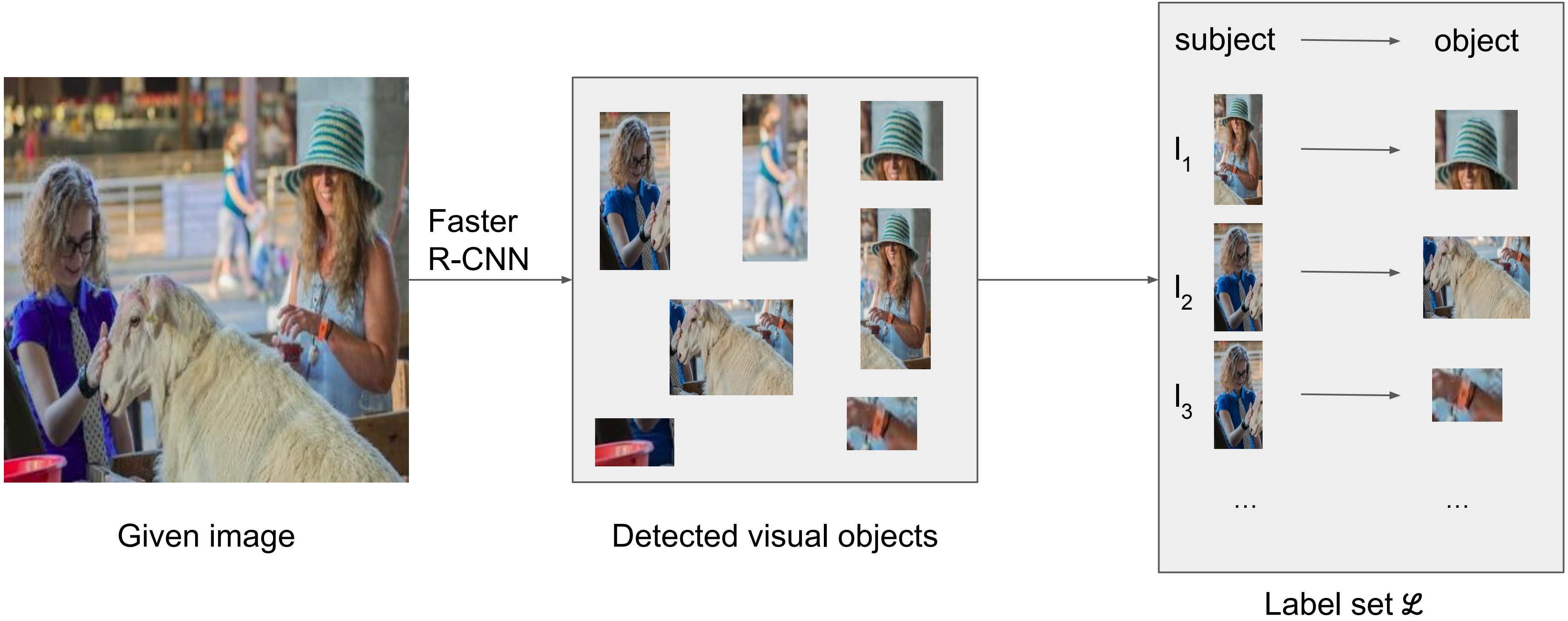
Visual relationship (VR) = <subject - predicate - object>

In this image : <woman - biting - apple>

In this work, to localize a VR we predict its :

subject bounding-box & object bounding-box

Problem Formulation : how a graph labelling problem : **what is a label set for an image**



Label set = all possible visual relationships in an image
= all possible ordered pairs of detected visual objects in an image

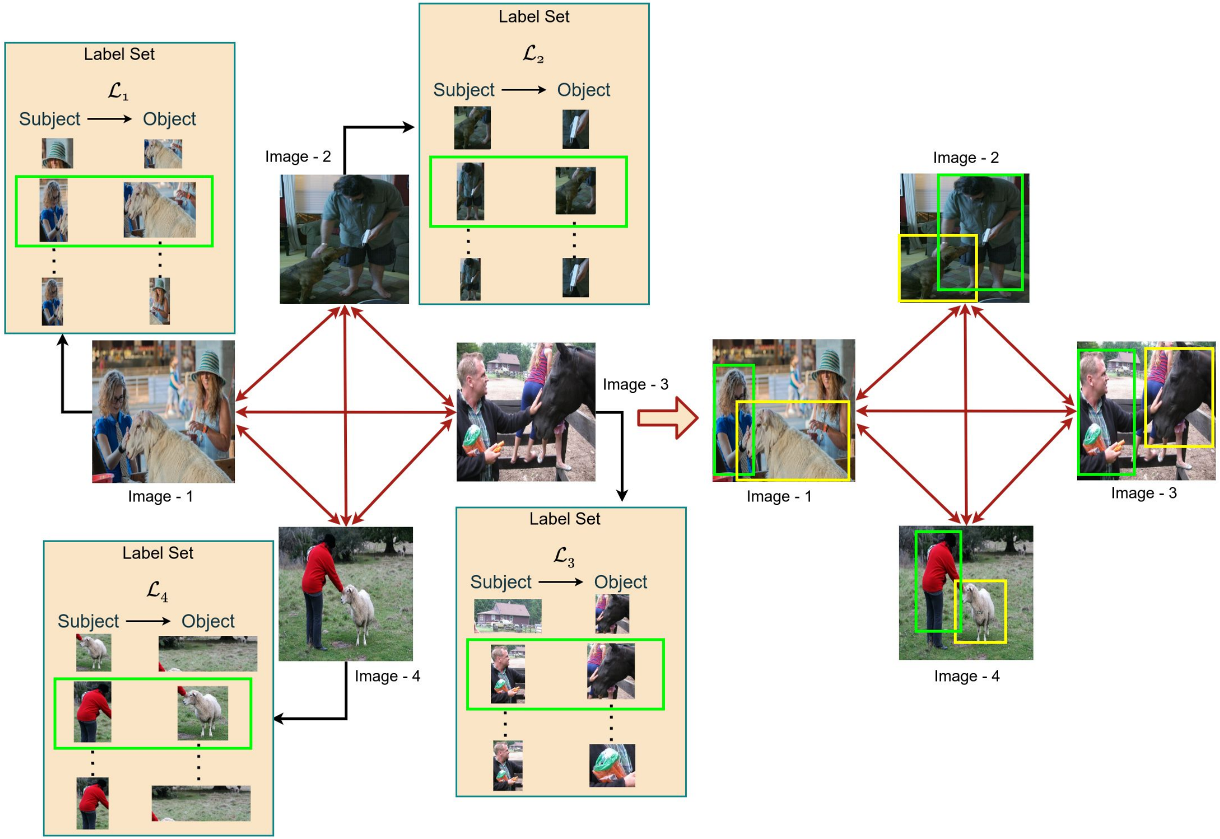
Problem Formulation : how a graph labelling problem : **what are nodes, edges and labels**

Each bag of image =
fully connected graph

Images in bag =
Graph vertices

Label set of image =
All possible VR, or
All possible subj-obj pairs

Objective =
Select 1 label (VR) for each image s.t. the selected labels have same predicate



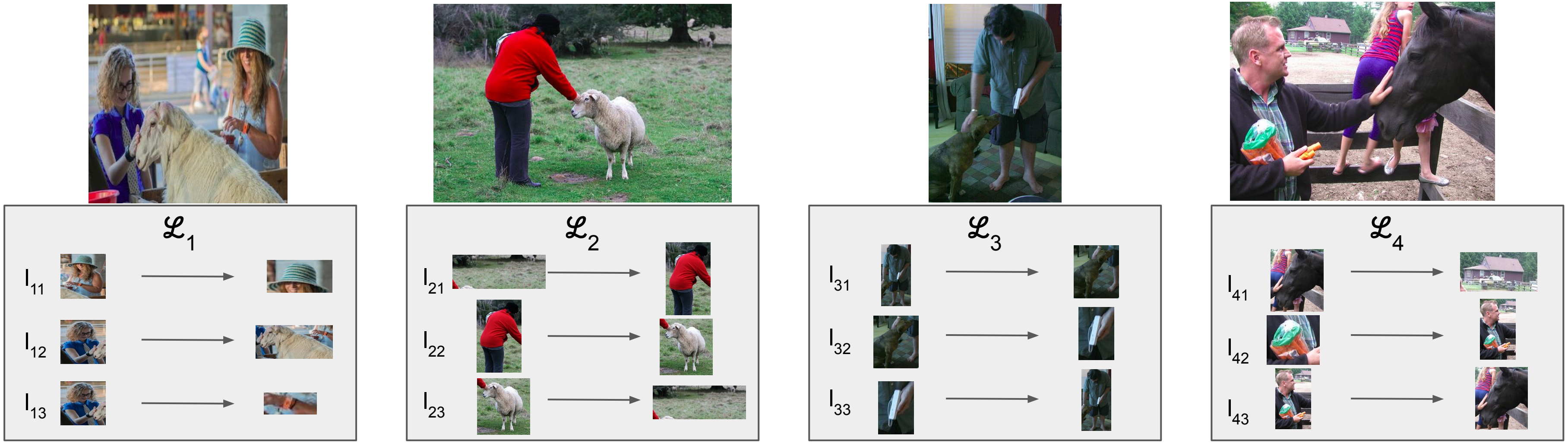
Bag of 4 images with common latent predicate = “petting”

Problem Formulation : how a graph labelling problem : **what is the cost function**

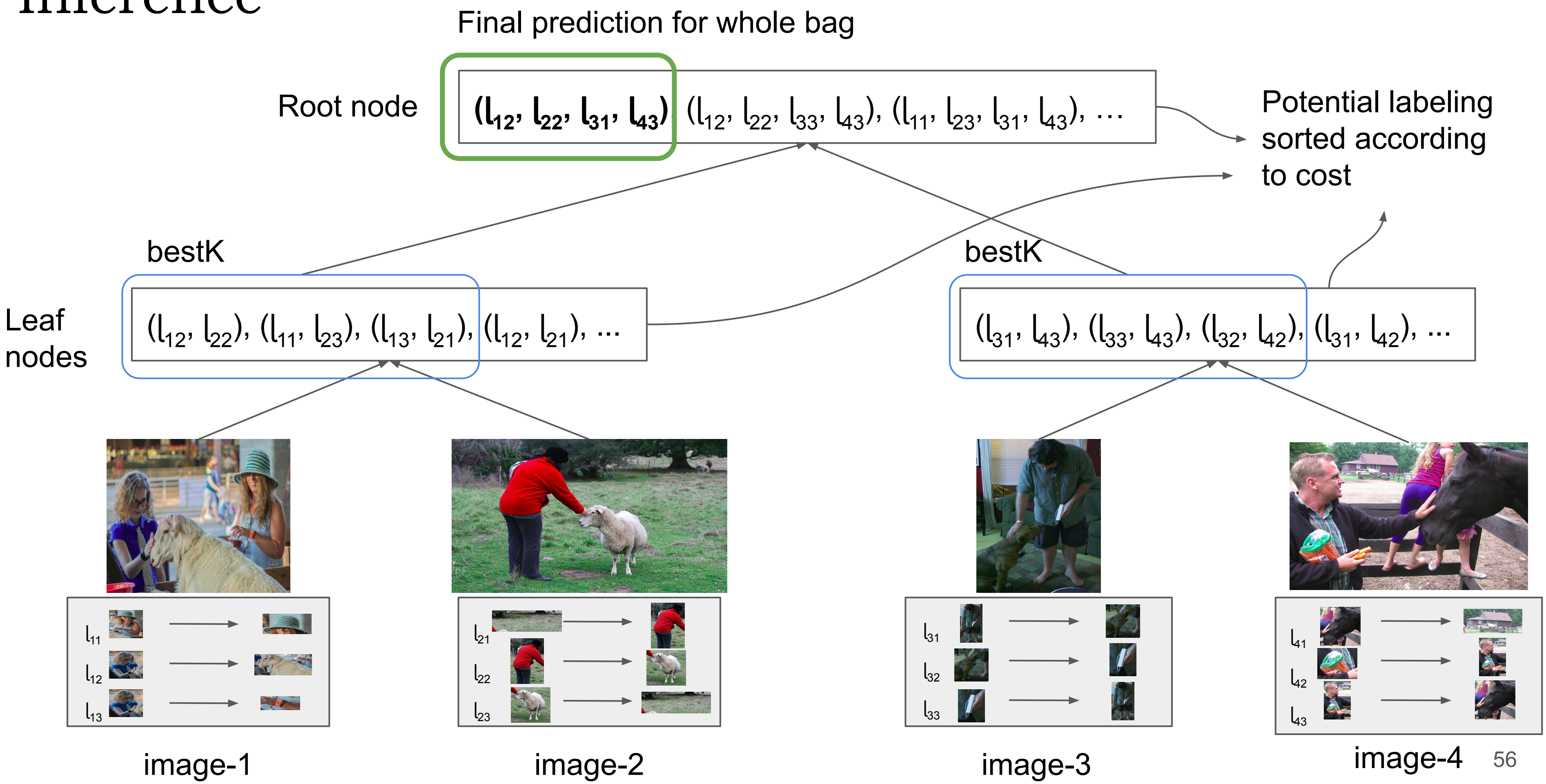
Optimization function : $$\Psi = \sum_{u=1}^{b} \left(\min_{t} \Psi_u(l_{ut}) + \sum_{v=1}^{b,u\neq v} \min_{t_1,t_2} \Psi_{uv}(l_{ut_1}, l_{vt_2}, \theta) \right)$$

Unary cost : $\Psi_u(l_{ut})$
Cost of assigning a label l_u to image u.
Considered uniform

Pairwise cost : $\Psi_{uv}(l_{ut_1}, l_{vt_2}, \theta)$
Cost of assigning labels l_{ut1} to image u and l_{vt2} to image v.
Lower when predicates of l_{ut1} and l_{vt2} are semantically similar



Inference


Final prediction for whole bag
Root node
$(l_{12}, l_{22}, l_{31}, l_{43})$, $(l_{12}, l_{22}, l_{33}, l_{43})$, $(l_{11}, l_{23}, l_{31}, l_{43})$, …
Potential labeling sorted according to cost
bestK
bestK
Leaf nodes
(l_{12}, l_{22}), (l_{11}, l_{23}), (l_{13}, l_{21}), (l_{12}, l_{21}), …
(l_{31}, l_{43}), (l_{33}, l_{43}), (l_{32}, l_{42}), (l_{31}, l_{42}), …
l_{11}
l_{12}
l_{13}
l_{21}
l_{22}
l_{23}
l_{31}
l_{32}
l_{33}
l_{41}
l_{42}
l_{43}
image-1
image-2
image-3
image-4

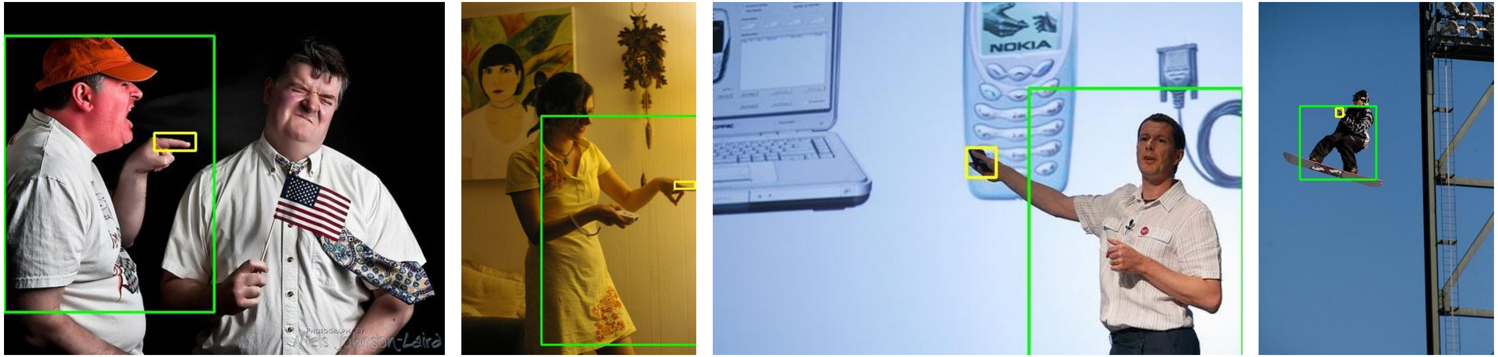
Where:
N_p = number of pairs in an episode
f_u, f_v = embeddings of visual relationship pairs and $u \neq v$
R_θ = visual relationship similarity function
Positive pairs: pairs sharing common predicate
Negative pairs: pairs sharing different predicate

Quantitative Results

Method →	Concat + Cosine			VtransE+ Cosine			Concat+ Rel. Net			Our Approach		
Supervision ↓	Bag Size			Bag Size			Bag Size			Bag Size		
	2	4	8	2	4	8	2	4	8	2	4	8
No supervision	72.16	70.86	76.85	73.34	74.20	82.56	75.61	74.02	76.38	78.99	76.12	84.07
Subject Fixed	76.82	78.66	**81.27**	80.37	**83.12**	83.58	**81.07**	**82.88**	**84.60**	83.90	**88.25**	86.67
Subject-Object in one image	**77.03**	**80.20**	79.42	**83.33**	82.40	**84.07**	79.29	81.69	81.45	**87.44**	84.46	**86.95**

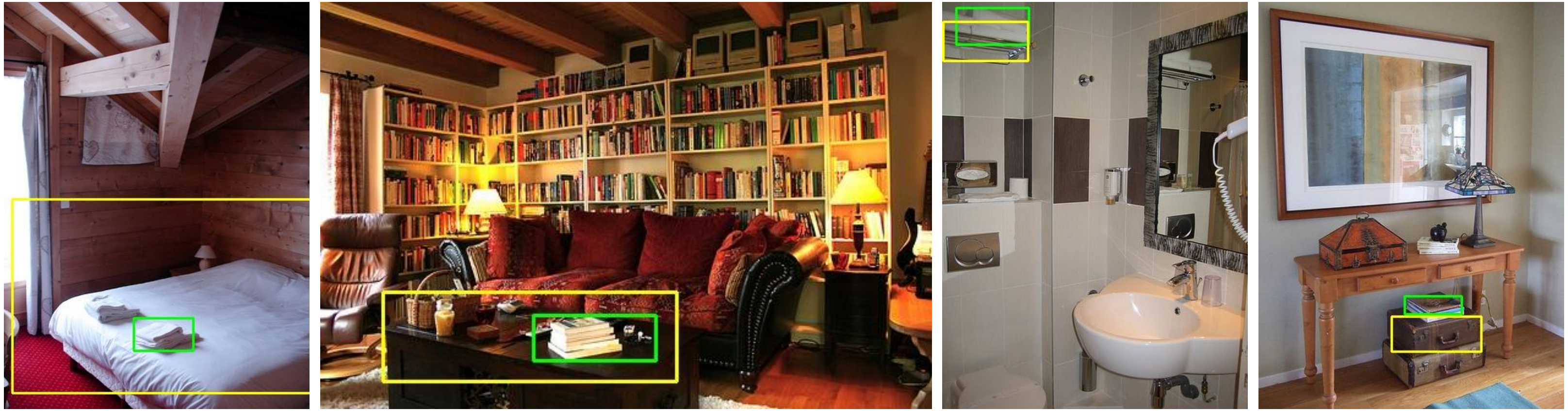
Table 3. **Effects of weak supervision on co-localization of relationships.** Here, we observe that just by giving a weak form of supervision, the visual relationship co-localization performance increases significantly for each ablation. The results correspond to VR-CorLoc %.

1:petting

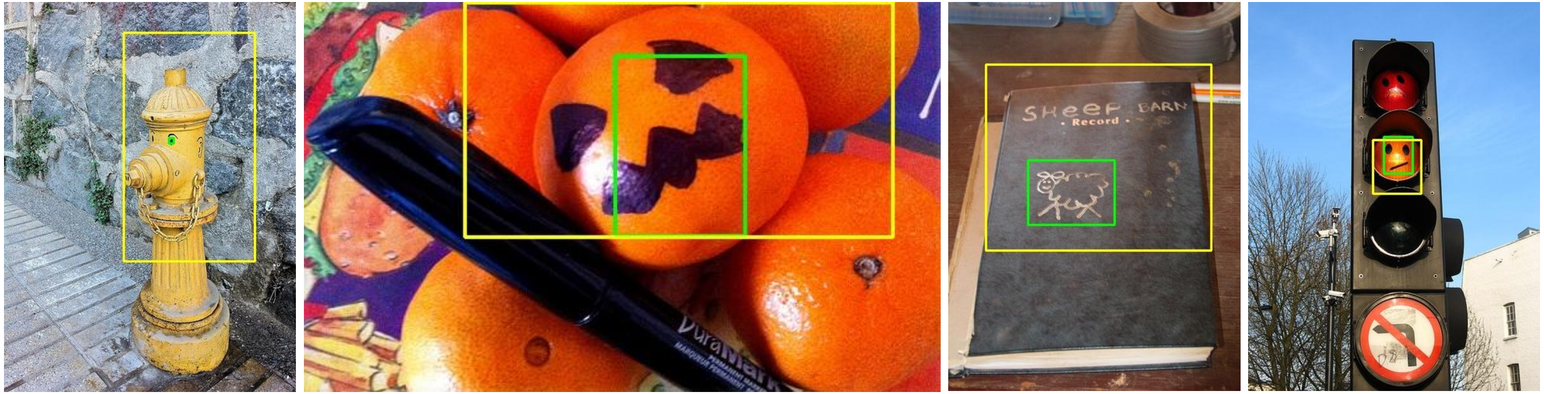


3:pointing

4:placed on

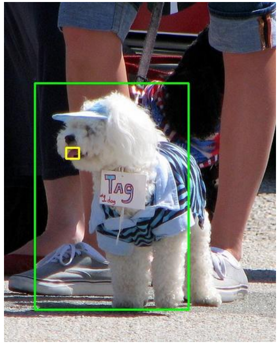


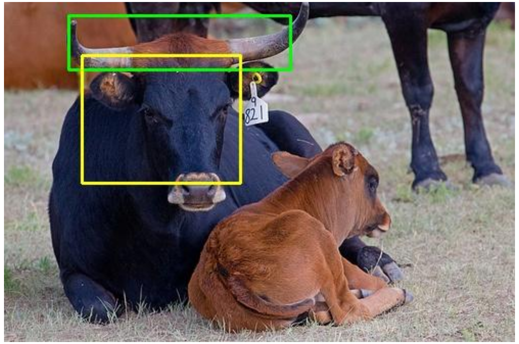
5:stacked on

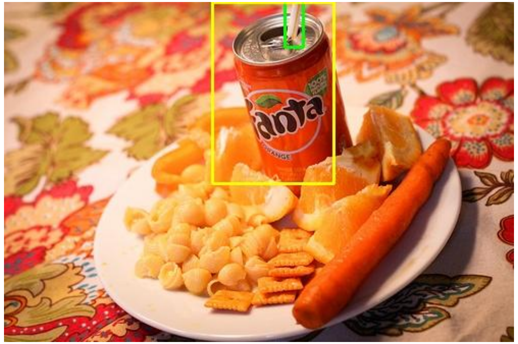


7:drawn on

8:sewn on

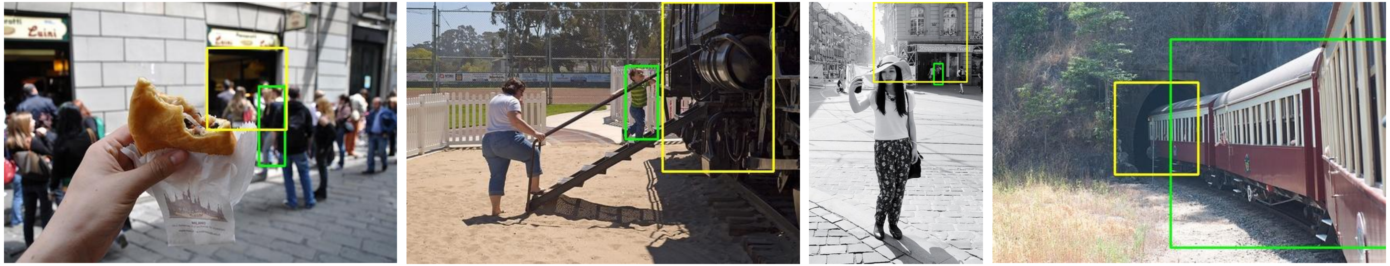




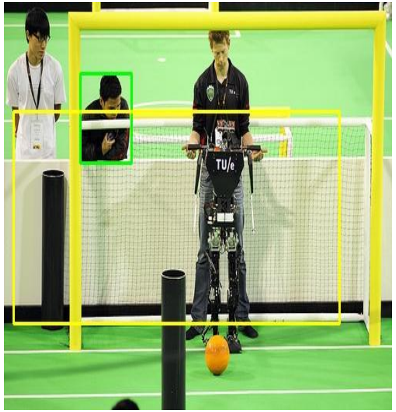



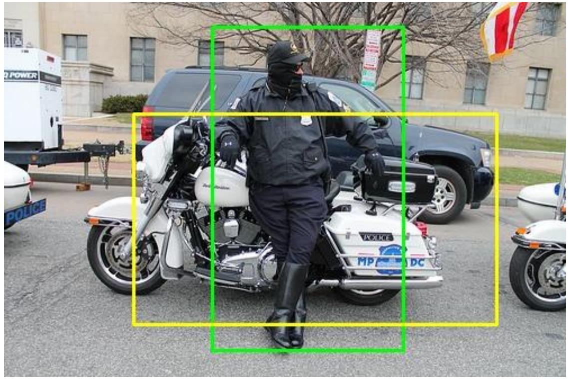
9:sticking out of

10:at bottom of

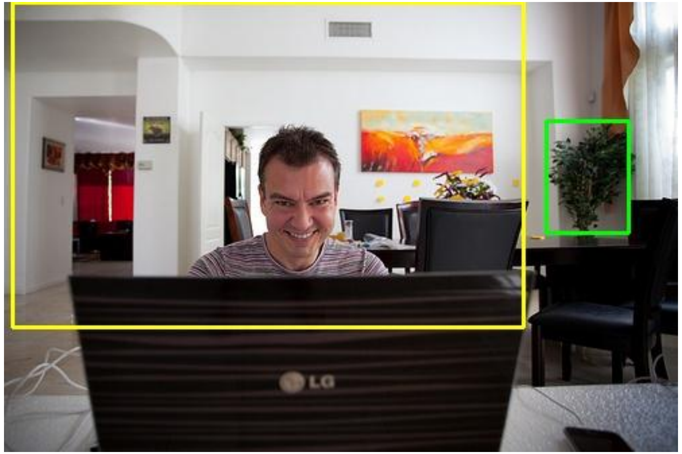


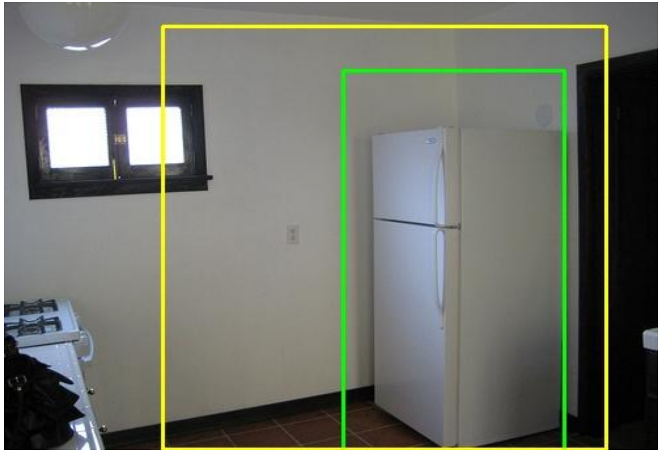
12:entering



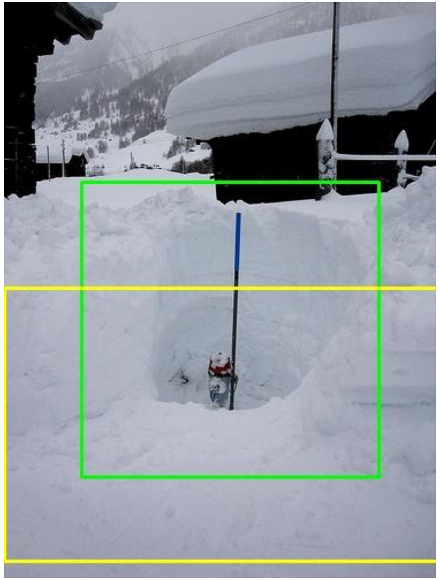


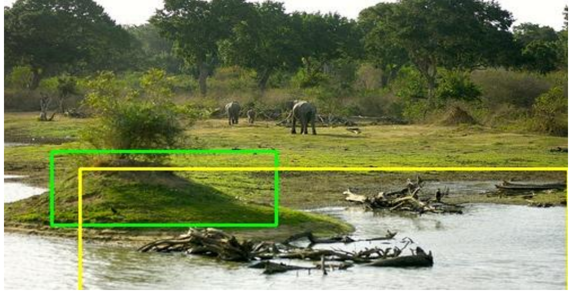

13:leaning on

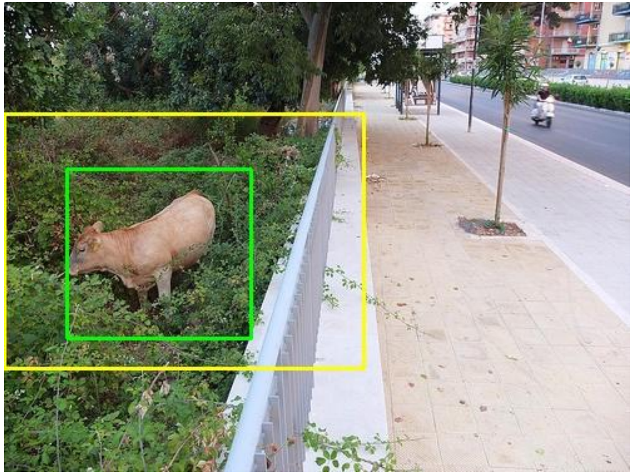




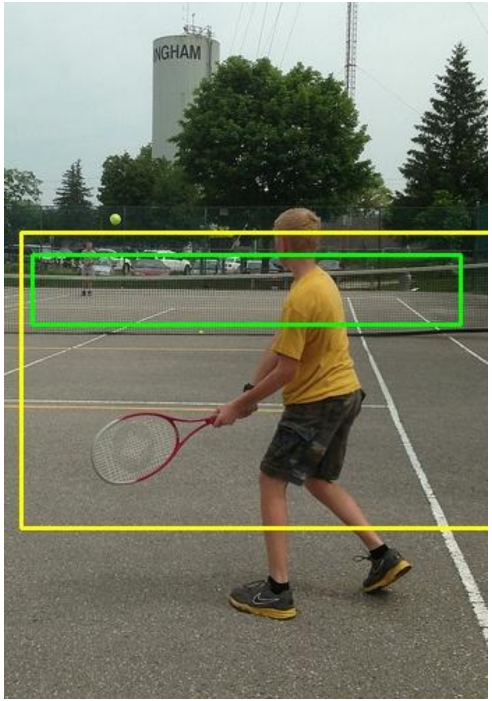
14:in corner of

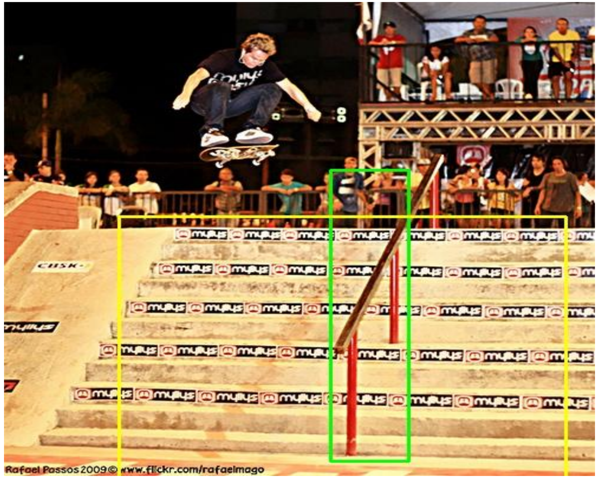


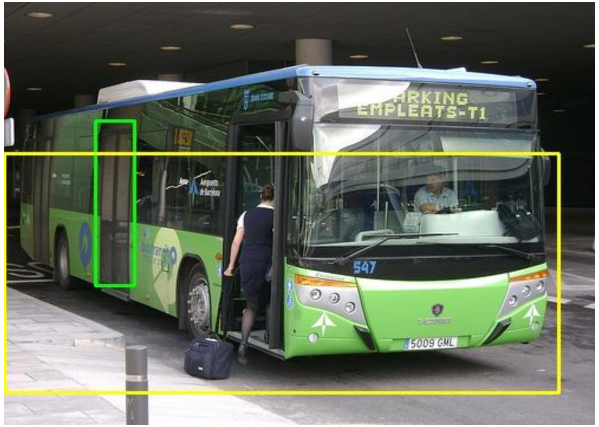




15: surrounded by







16:in center of

Episodic training with binary logistic regression loss

For positive pairs: $$L^p = \frac{1}{N_p} \sum_{(f_u, f_v)} \Big(log(1 + exp(-R_\Theta(f_u, f_v)) \Big)$$

For negative pairs: $$L^p = \frac{1}{N_p} \sum_{(f_u, f_v)} \Big(log(1 + exp(R_\Theta(f_u, f_v)) \Big)$$

Where:

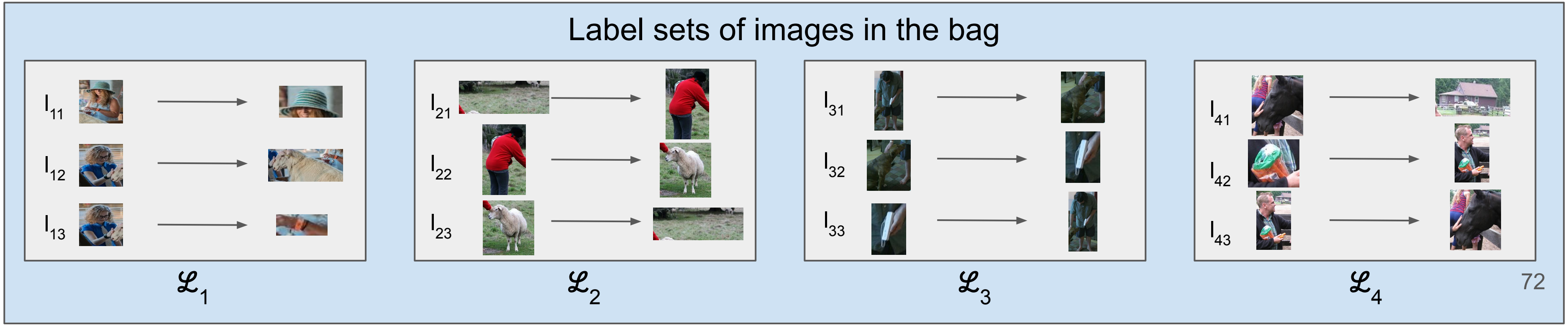
N_p = number of pairs in an episode

f_u, f_v = embeddings of visual relationship pairs and u ≠ v

R_θ = visual relationship similarity function

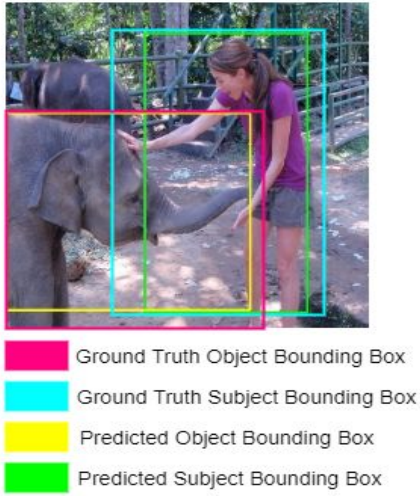
Positive pairs: pairs sharing common predicate

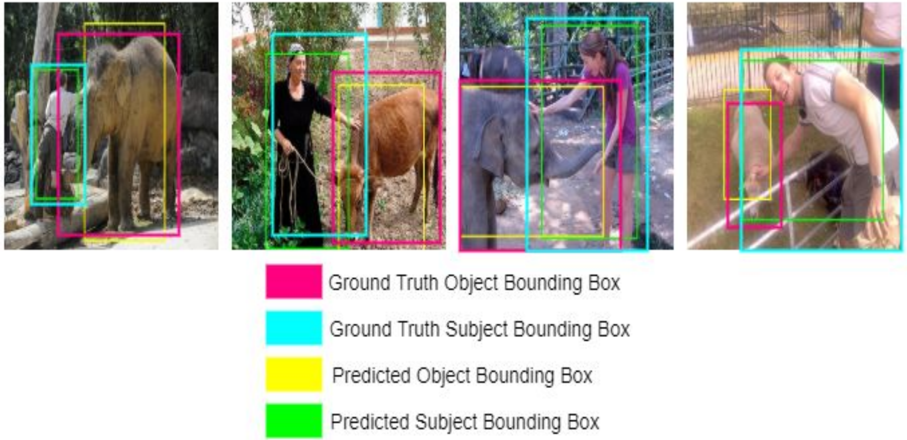
Negative pairs: pairs sharing different predicate



Supple slides

Metrics





VR-CorLoc

Fraction of test images for which visual subject-object pairs are correctly localized.

Bag-CorLoc

Fraction of the total number of bags for which the visual subject-object pairs are correctly localized for all of its images.