Leibniz Universität Hannover

Retina-Net: Which Relationships Humans Pay More Attention to

Jing Xu

Erstprüfer: Prof. Bodo Rosenhahn

Betreuer: Yuren Cong
Institut für Informationsverarbeitung

Outline

Introduction

Our Method

Dataset and Evaluation Metrics

Experimental Results

Conclusions

Outline

Introduction

- Background
- Related Works
- Motivation

Our Method

Dataset and Evaluation Metrics

Experimental Results

Conclusions

Introduction

Scene Graph Generation/Visual Relationship Detection:

understand the relationship between any two of objects in an image/video.

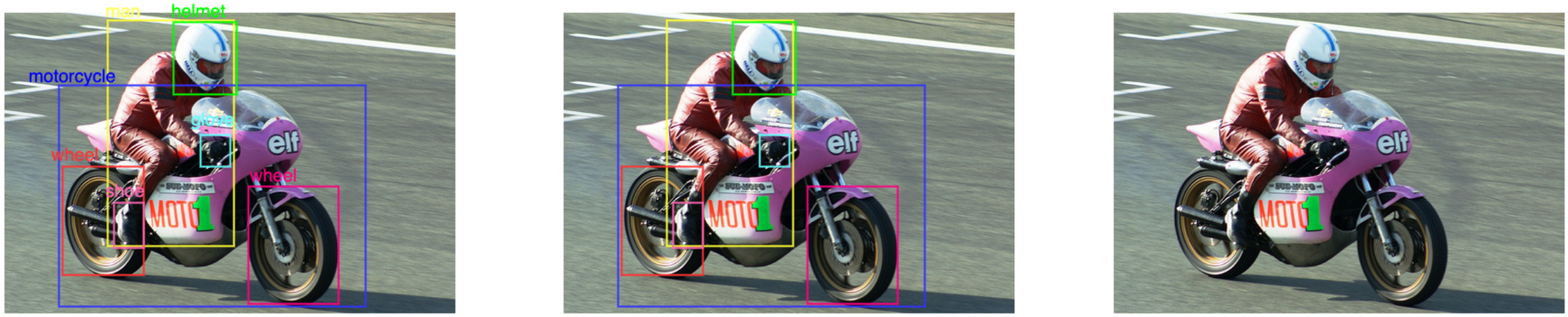
Evaluation settings :

Predicate Classification (**PredCLS**)

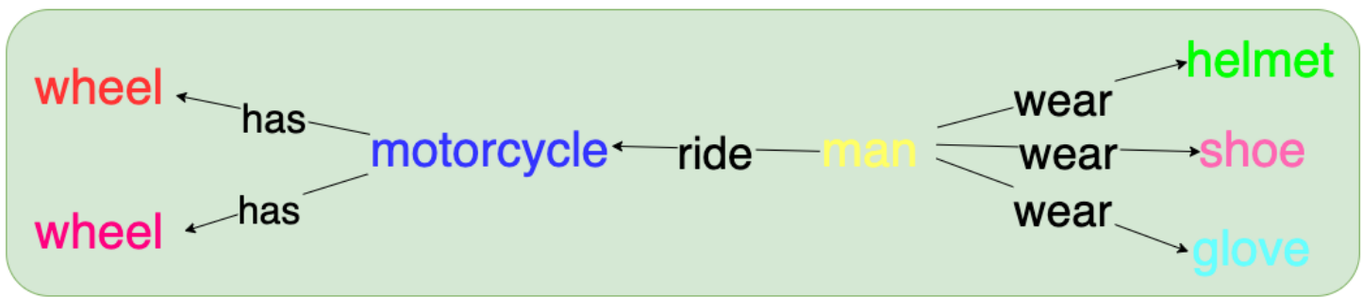
Scene Graph Classification (**SGCLS**)

Scene Graph Detection (**SGDET**)

Input:



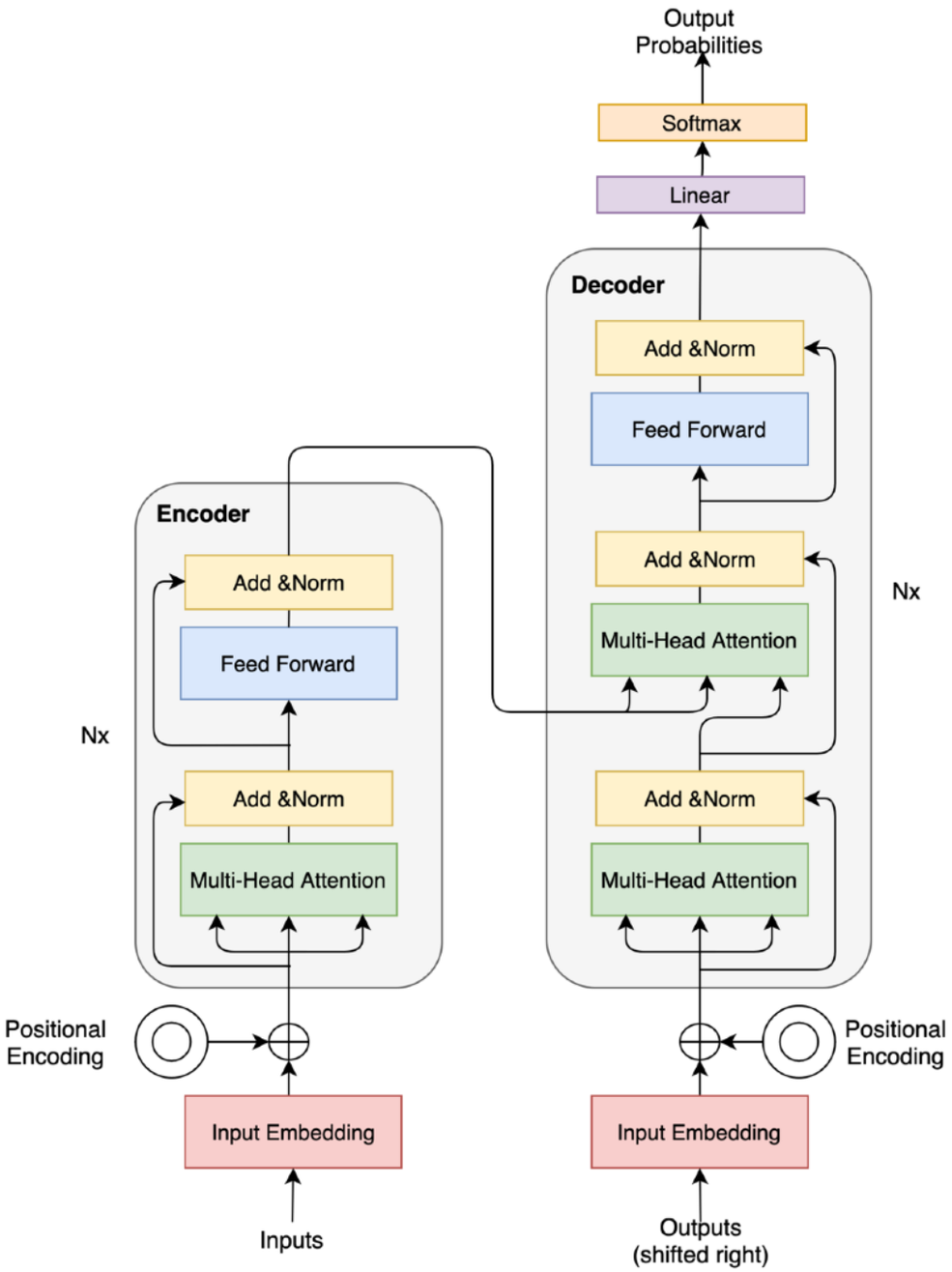
Output:



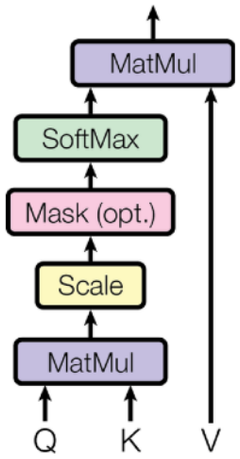
Background

Transformer:

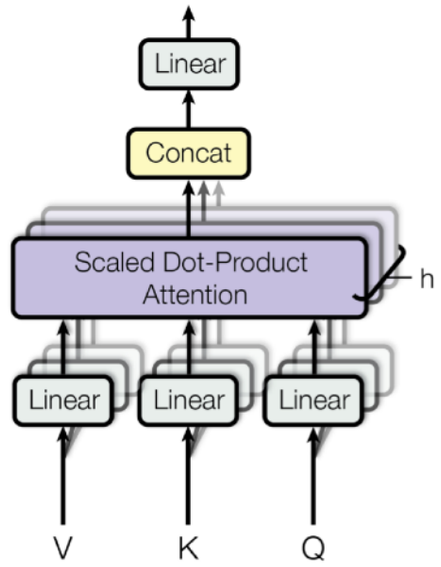
A sequence to sequence architecture base on attention-mechanism.



Scaled Dot-Product Attention



Multi-Head Attention

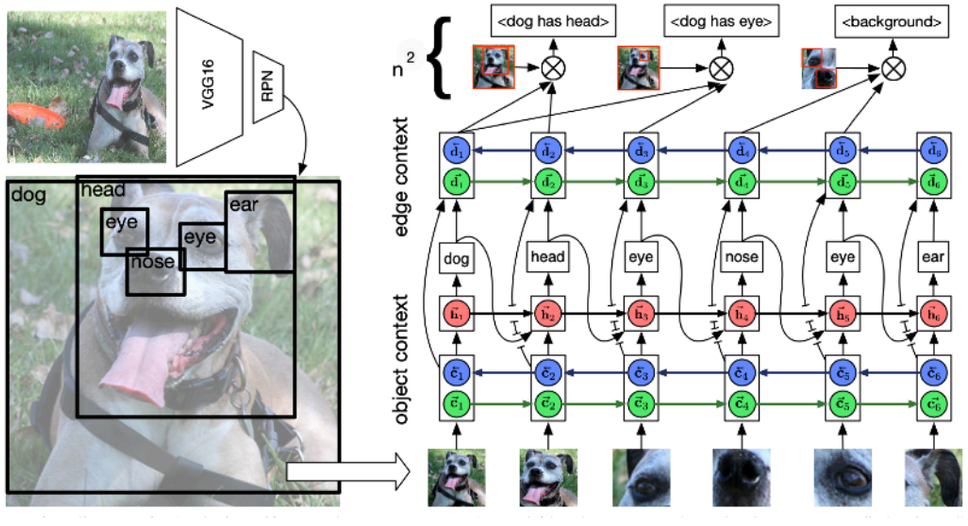


$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_K}})V$$

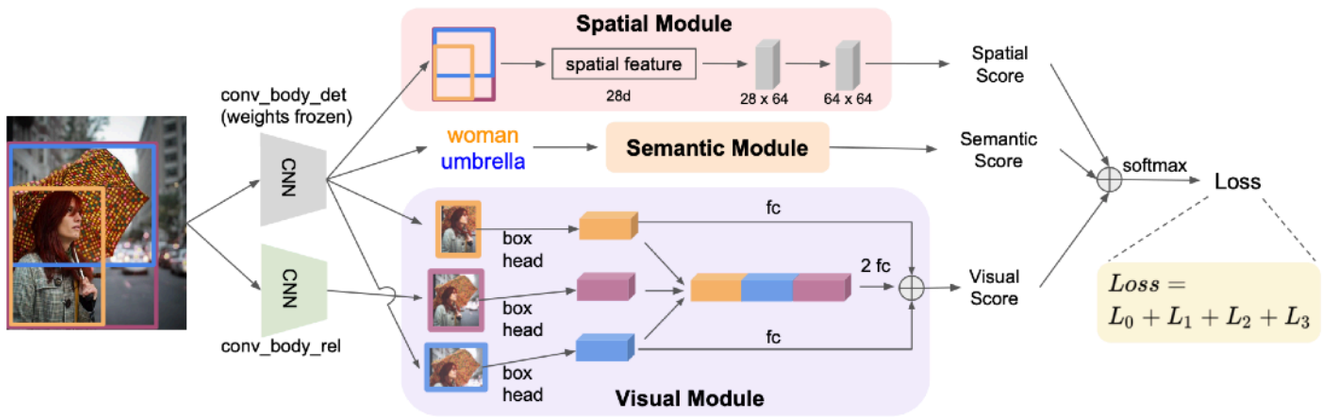
Jing Xu
xujing@tnt.uni-hannover.de

Related work

Neural Motifs: Scene Graph Parsing with Global Context : They devise an efficient mechanism (LSTMs) to encode the global context that can directly inform the local predictors in modelling scene graphs.



RelDN: Graphical Contrastive Losses for Scene Graph Parsing : They propose a set of Graphical Contrastive Losses to solve Entity Instance Confusion and Proximal Relationship Ambiguity problems.

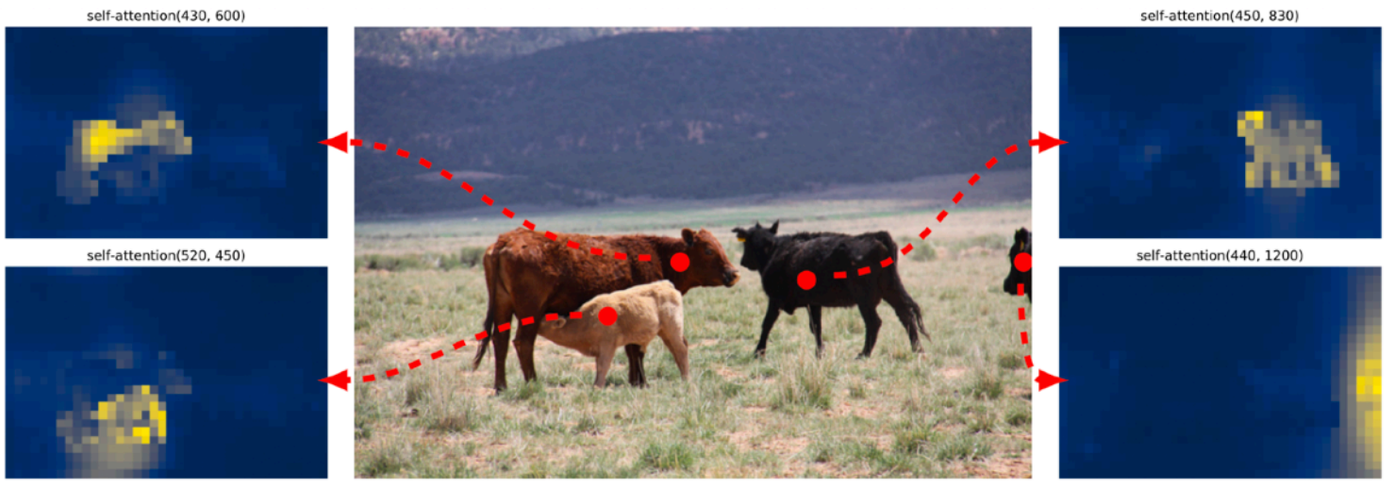


Related work

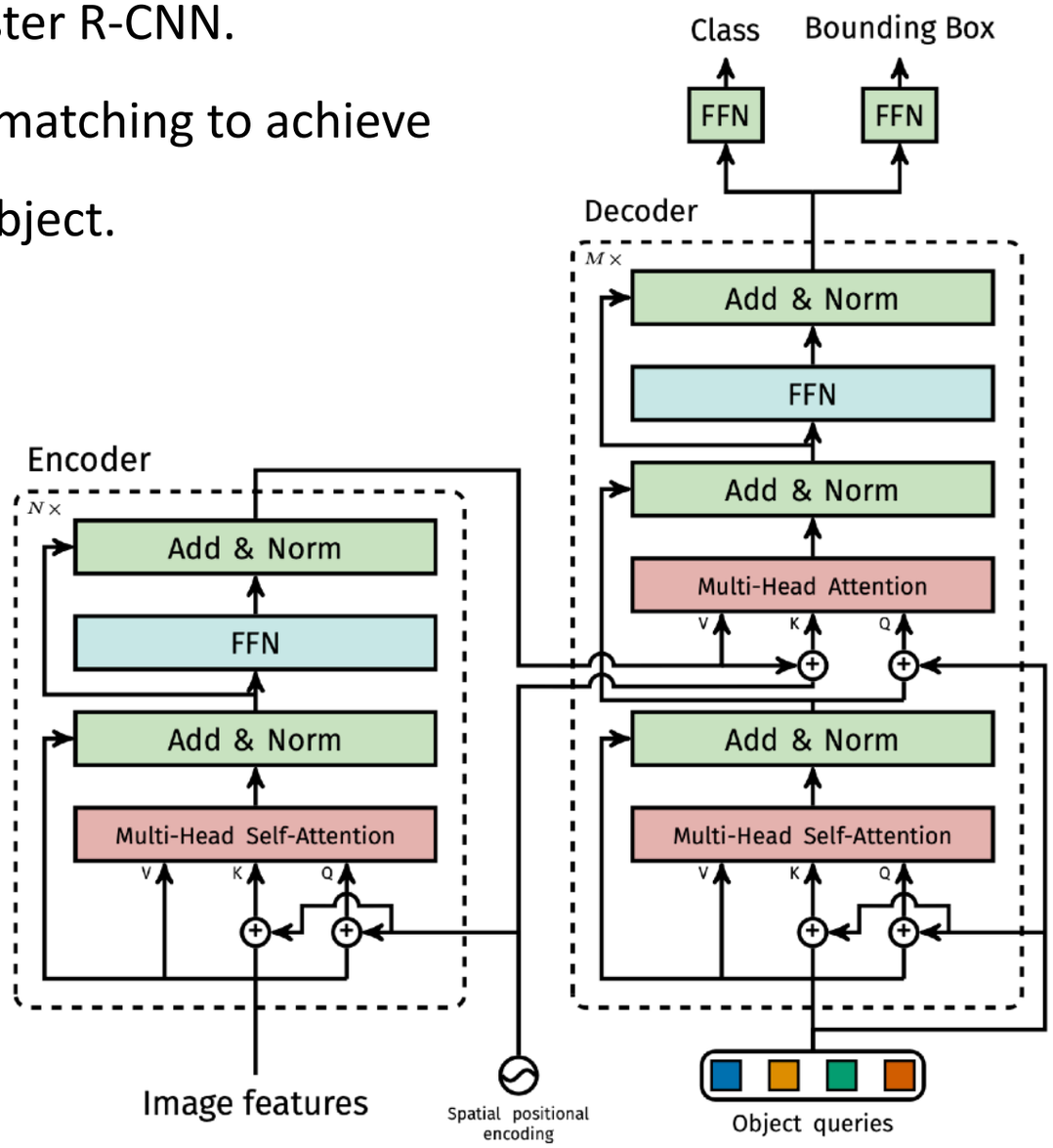
DETR: End-to-End Object Detection with Transformers:

They adopt a transformer encoder-decoder architecture in object detection problem.

1. It achieves an accuracy comparable to or even surpassing Faster R-CNN.
2. It used a fixed learnable query in each image and a bipartite matching to achieve position-independent and unique bounding box for each object.
3. Each object can be visualised in the attention map well.



The attention map of the DETR



Architecture of DETR's transformer

Motivation

Motivation

- Proposed a solution for visual relation detection through the transformer structure.
- Design an object query for predicate classification and scene graph classification.
- We hope to obtain some information that is conducive to solve VRD problems by attention mechanism.
- The most challenging is to predict the correct predicate, which describes the relation between two objects.

Outline

Introduction

Our Method

- Pixel-based Attention
- Retina-Net

Dataset and Evaluation Metrics

Experimental Results

Conclusions

Pixel-based Attention

Idea

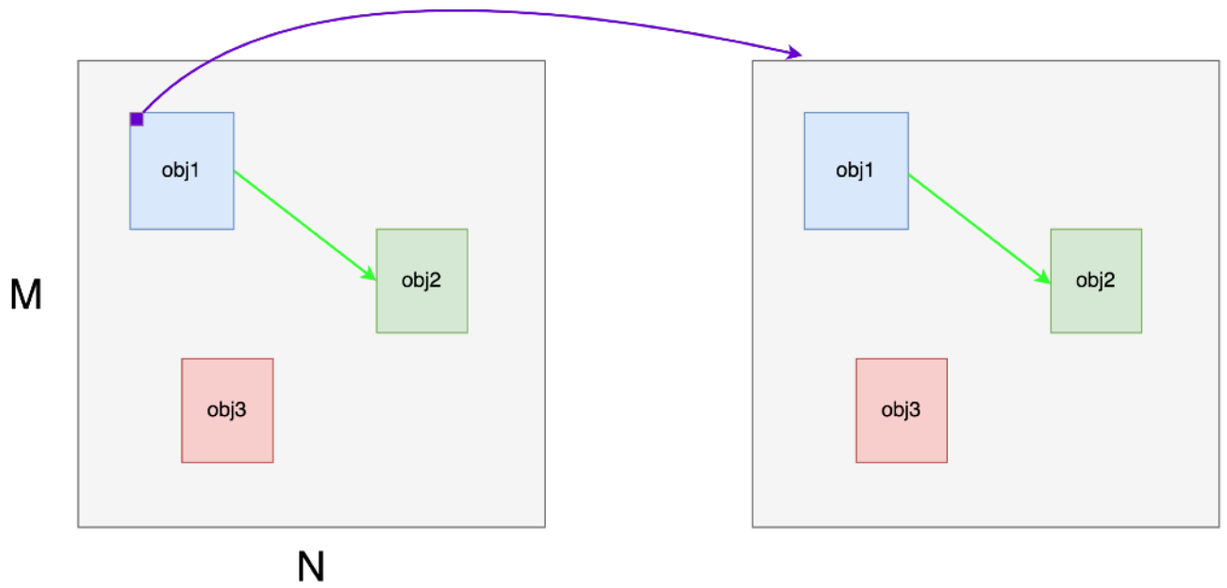


Fig1: The attention between the first pixel of object$_1$ and the whole image.

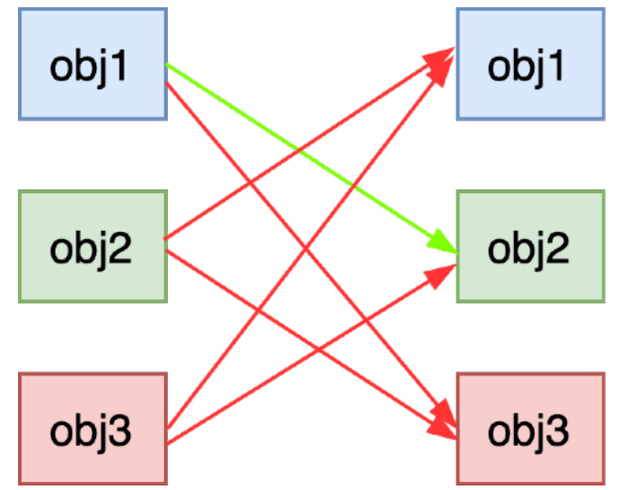


Fig2: The relationship between objects.

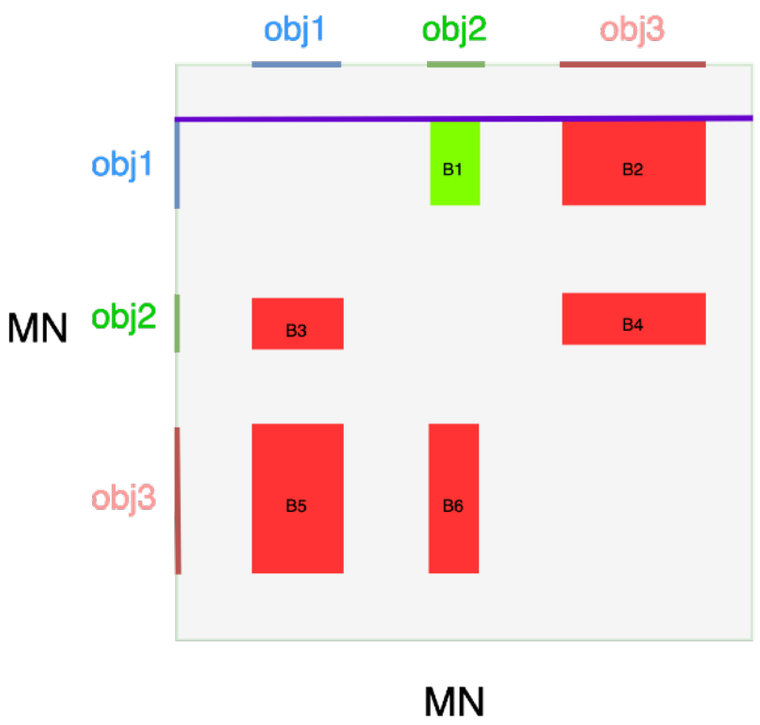


Fig3: The attention map between pixels of objects.

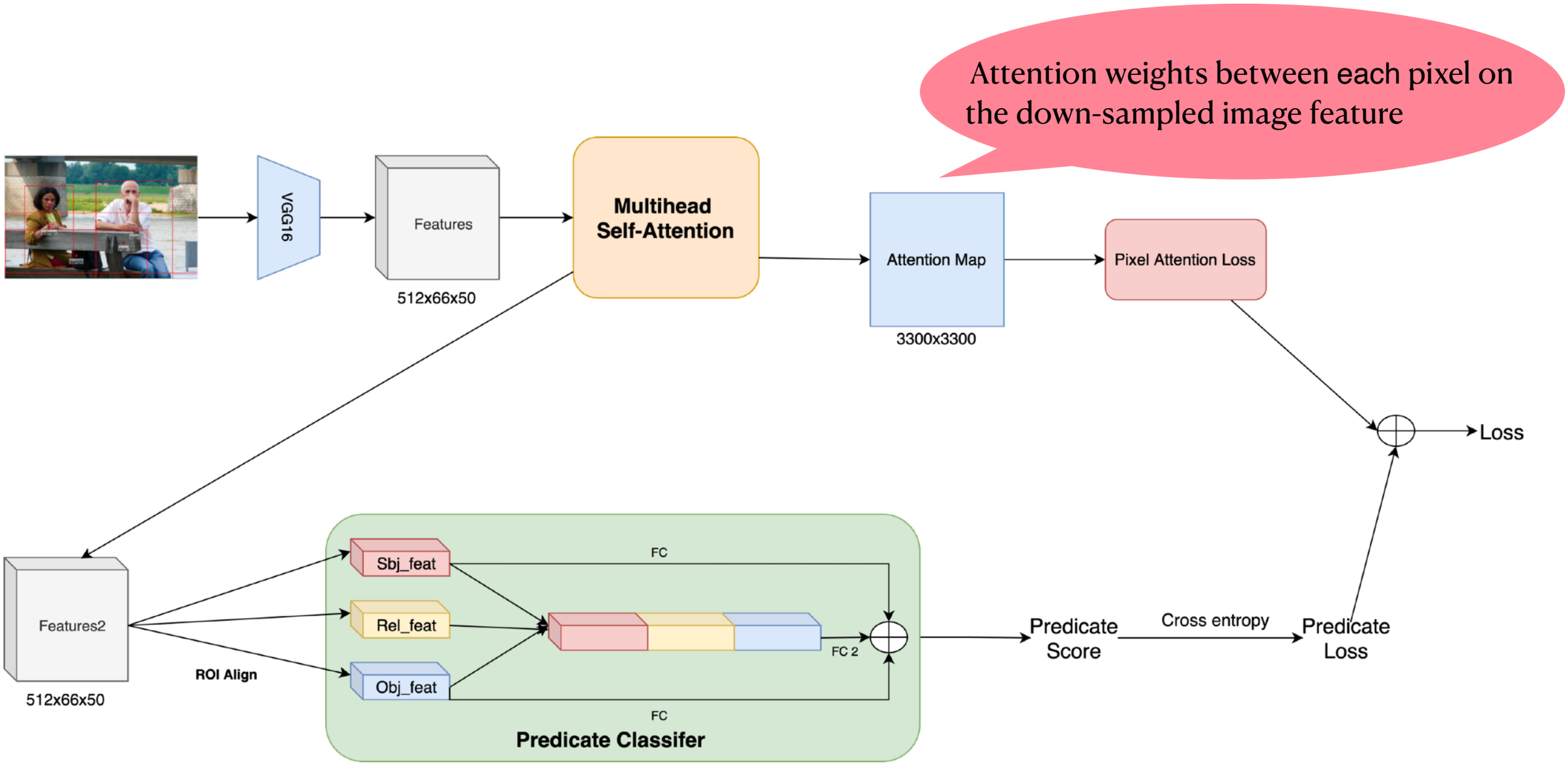
: The attention between the first pixel of object 1 and the whole image

: The attention of the ground truth relation pair

: The attention of the no relation pair

Pixel-based Attention

Implementation



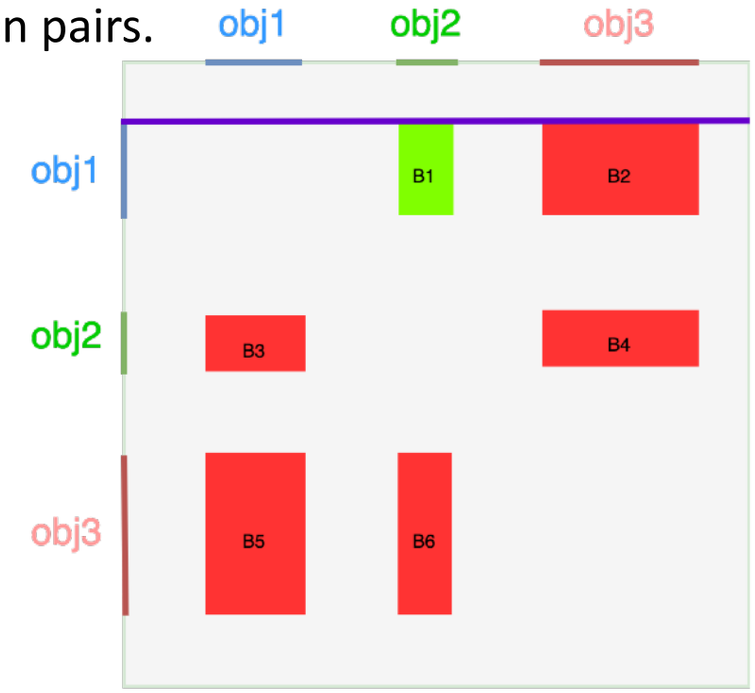
Pixel-based Attention

Pixel Attention Loss

$$loss_{attention} = max(0, \frac{1}{m}\sum_{m} Att_j^{no\_rel} - \frac{1}{n}\sum_{n} Att_i^{rel} + M)$$

Where:

- $Att_j^{no\_rel}$: the attention weight of the pixel j of no-relation pairs.
- Att_i^{rel}: the attention weight of the pixel i of ground true relation pairs.
- M: margin.
- m: the sum of $Att_j^{no\_rel}$.
- n: the sum of Att_i^{rel}.



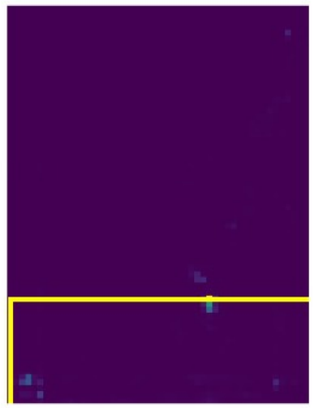
Shortcoming :

There is serious overlap between the relation pair and the no relation pair.

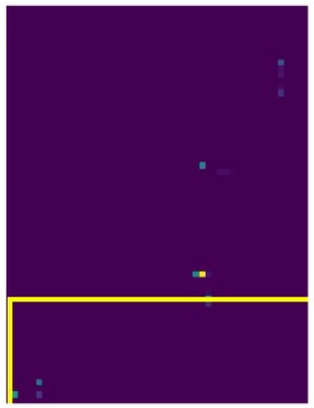
Pixel-based Attention

Visualised results

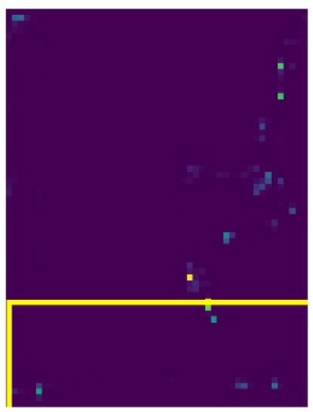
Image: dog sit on beach



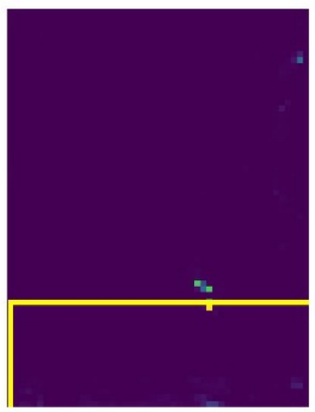
Pixel attention map (1)



Pixel attention map (2)

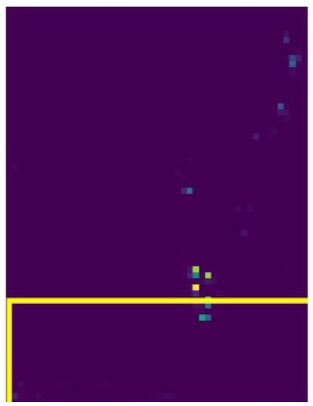


Pixel attention map (3)

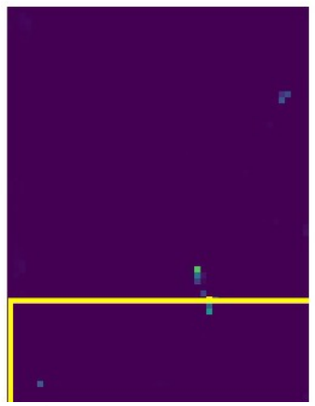


Pixel attention map (4)

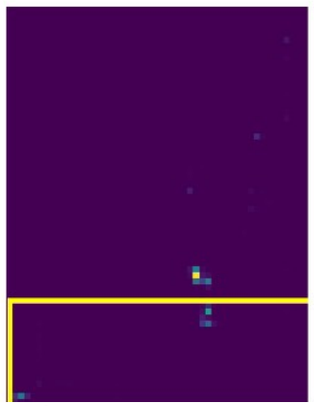
Pixel attention map (5)



Pixel attention map (6)



Pixel attention map (7)



Pixel attention map (8)

Relation pair: <dog sit on beach>
Red box: subject <dog>
Yellow box: object <beach>

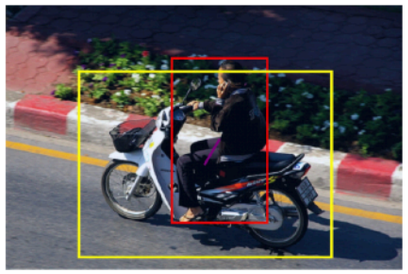
Result:

The subject<dog> don't Pay attention to the object <beach>, it focus on itself more.

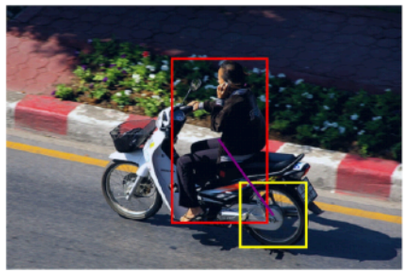
Pixel-based Attention

Result Analysis

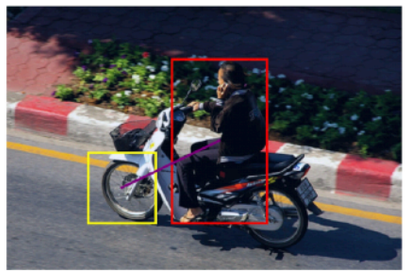
Relation pair



(a) $< man, ride, bike >$

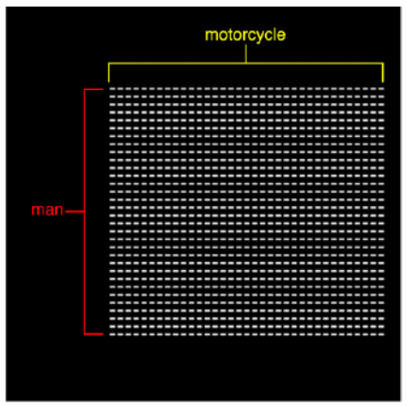


(b) $< man, \varnothing, wheel_1 >$

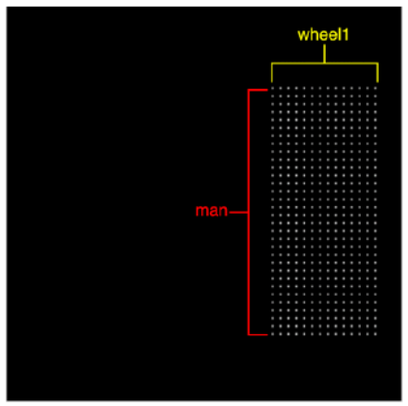


(c) $< man, \varnothing, wheel_2 >$

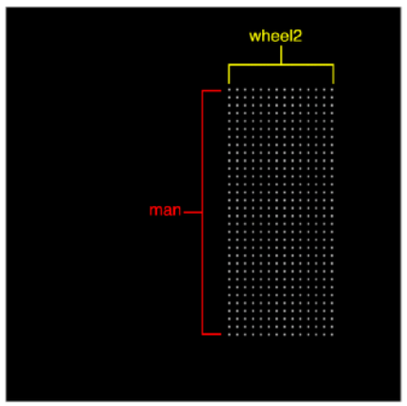
Position of a relation pair in attention map



(d) $\forall Attention_{p^i_{man} \rightarrow p^j_{bike}}$

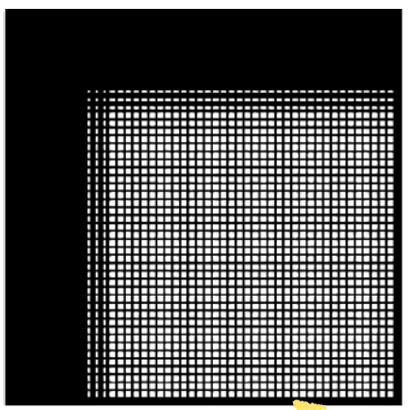


(e) $\forall Attention_{p^i_{man} \rightarrow p^j_{wheel_1}}$

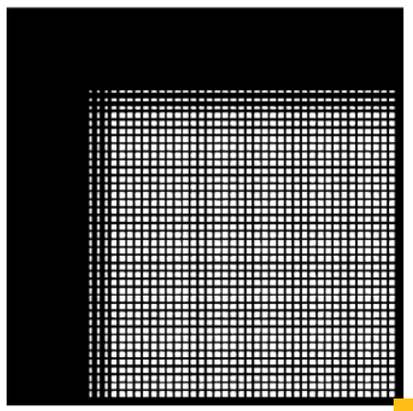


(f) $\forall Attention_{p^i_{man} \rightarrow p^j_{wheel_2}}$

(g) $\forall Att^i_{no\_rel}$



(h) $\forall Att^i_{rel}$



(i) $\forall Att^i_{rel} \cap \forall Att^i_{no\_rel}$

Overlap

$$loss_{attention} = max(0, \frac{1}{m}\sum_{m} Att_j^{no\_rel} - \frac{1}{n}\sum_{n} Att_i^{rel} + M)$$

Reason:

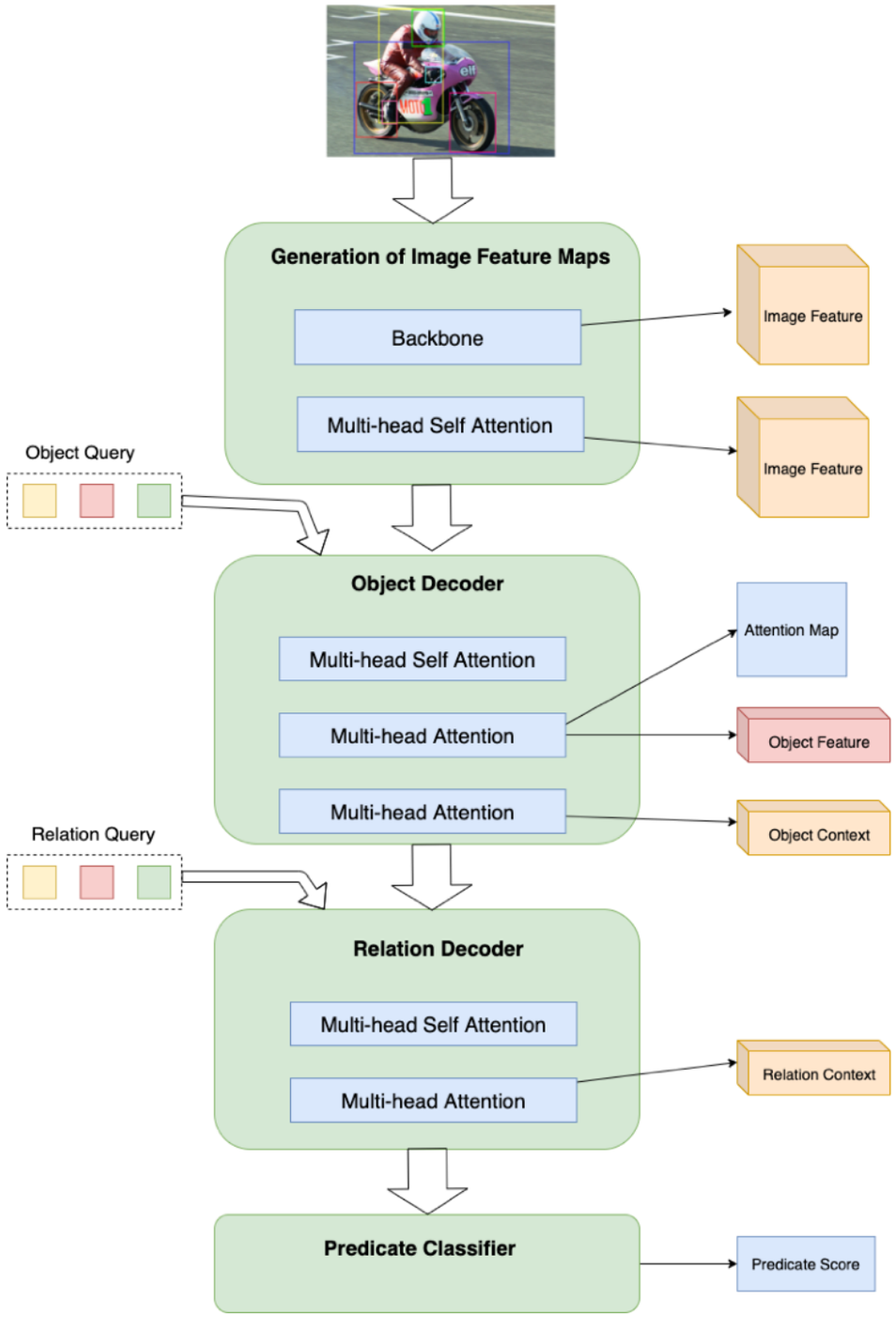
The overlap of attention weights between ground truth relation pairs and no relation pairs is common, causing attention loss to not work.

Solution:

Avoid using pixel-based attention and we proposed Retina Net based transformer structure.

Retina Net

Implementation



There are four components in Retina Net:

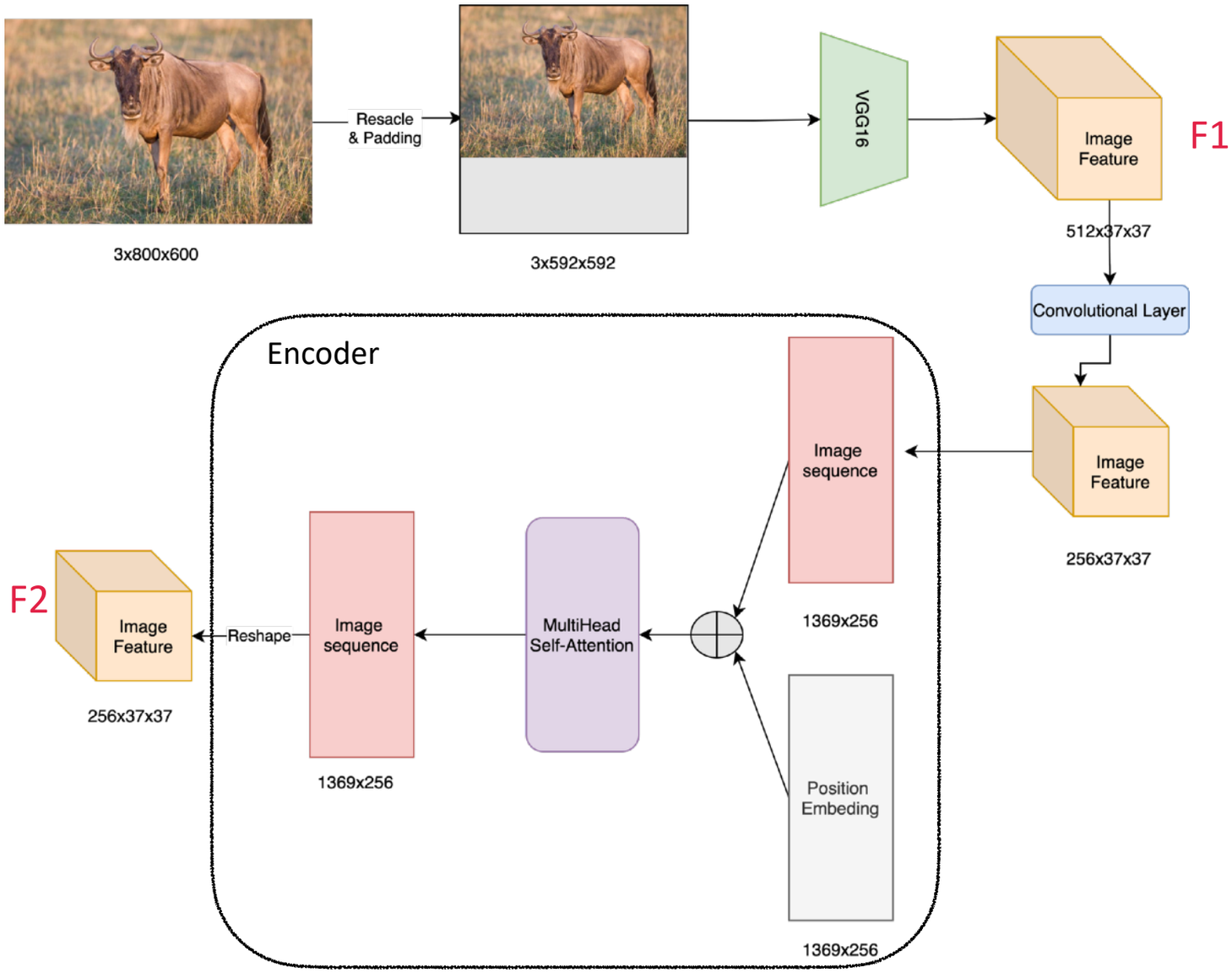
- Generation of Image Feature Maps
  generate visual features of the images.
- Object Decoder
  obtain object features and the context between each entity.
- Relation Decoder
  obtain the context between entities and relations.
- Predicate Classifier
  predict predicates.

Innovation:

- Propose a model based on the transformer structure to solve the VRD problem.
- Design a new object query for PredCLS and SGCLS.
- Represent the global context via Multi-head Attention module.

Retina Net

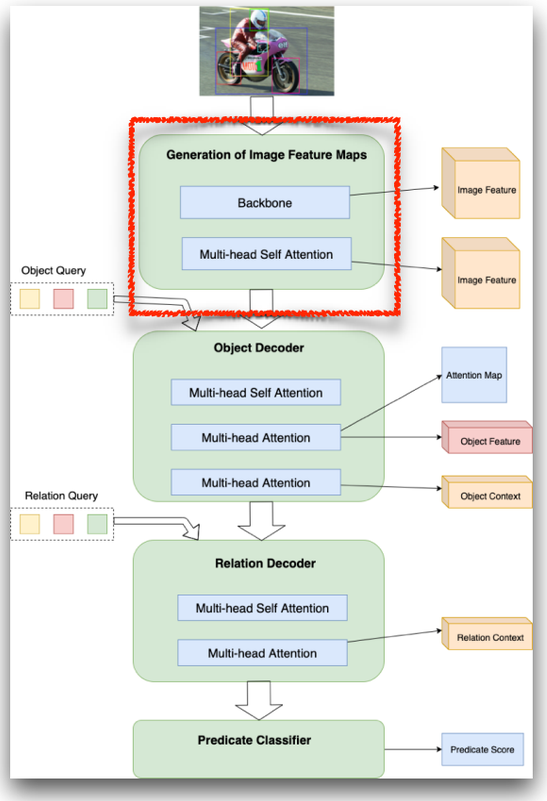
Generation of Image Feature Maps



Two visual image features:

F1: obtained from VGG16

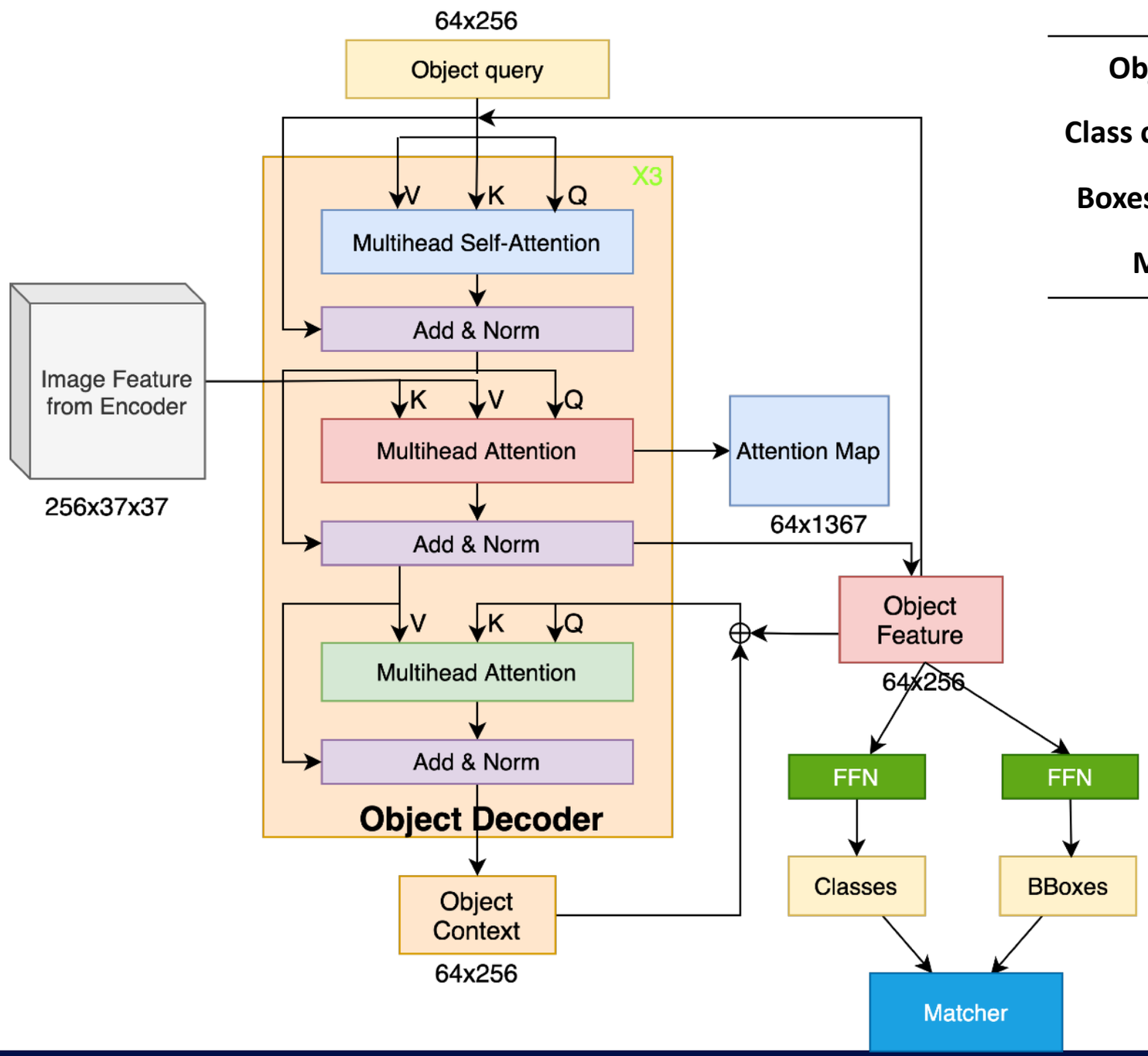
F2: pixel-wise interactive



Retina Net

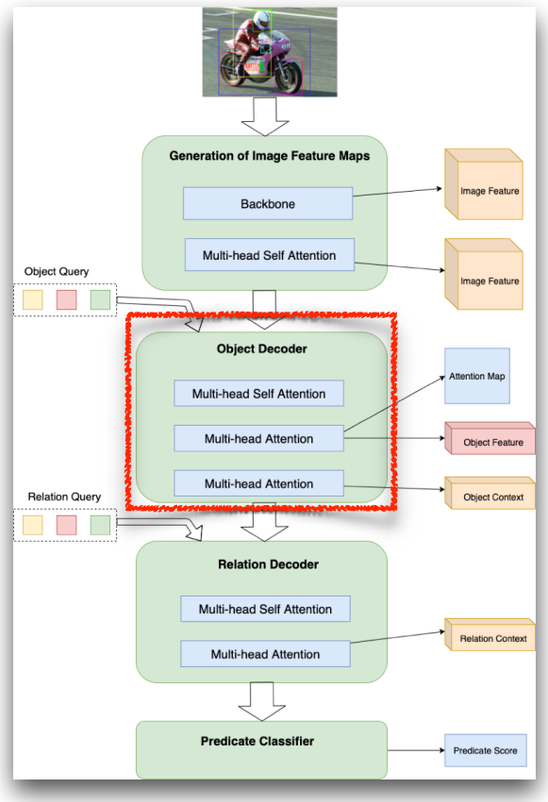
Object Decoder

Architechture:



	PredCLS	SGCLS	SGDET
Objet query	Box query	Box query	Learnable query
Class classification	No	Yes	Yes
Boxes prediction	No	No	Yes
Matcher	No	No	Yes

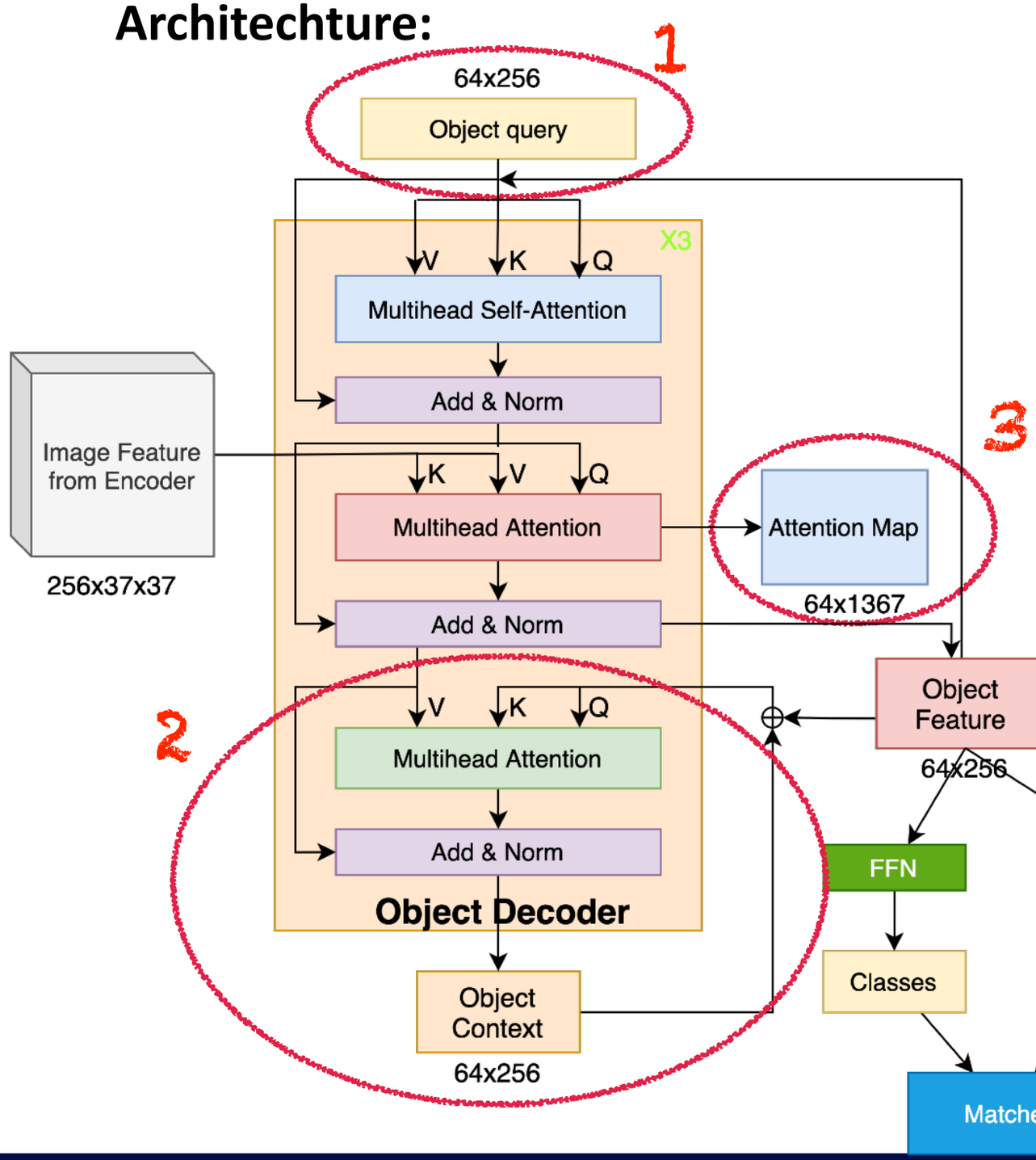
The setting of the object decoder.



Retina Net

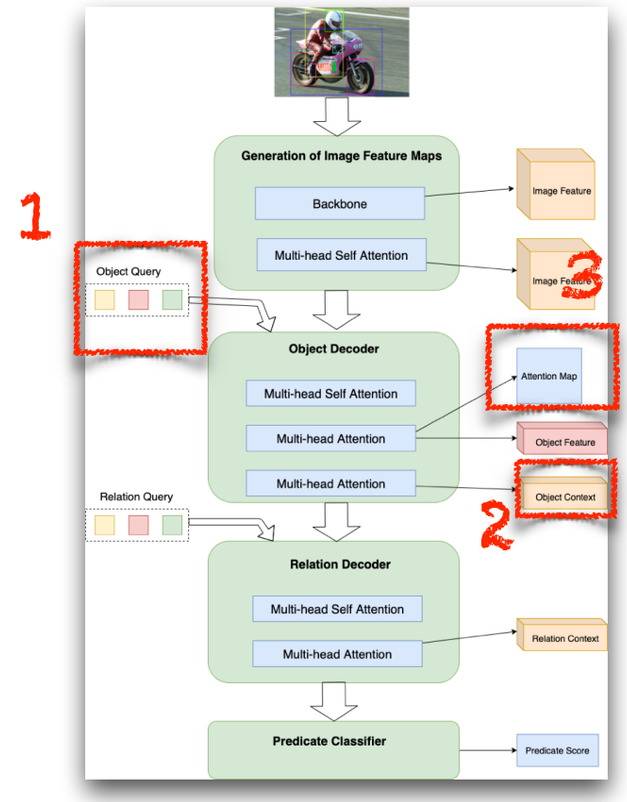
Object Decoder

Architechture:



	PredCLS	SGCLS	SGDET
Objet query	Box query	Box query	Learnable query
Class classification	No	Yes	Yes
Boxes prediction	No	No	Yes
Matcher	No	No	Yes

The setting of the object decoder.

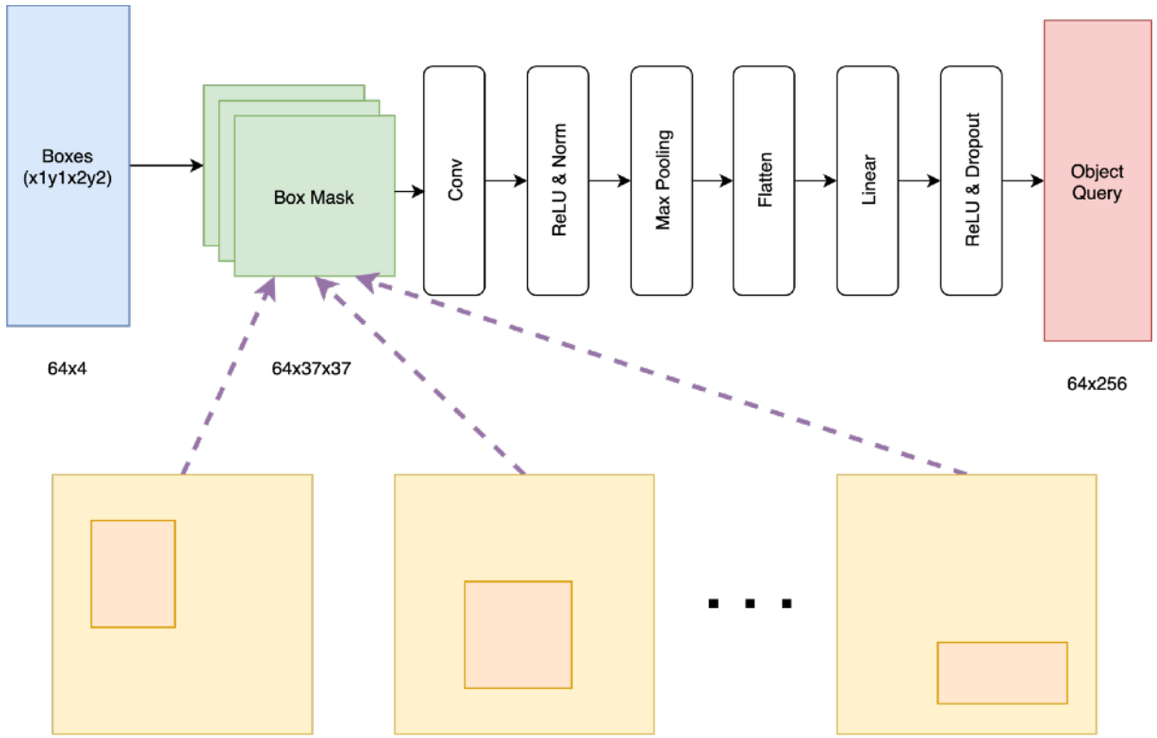


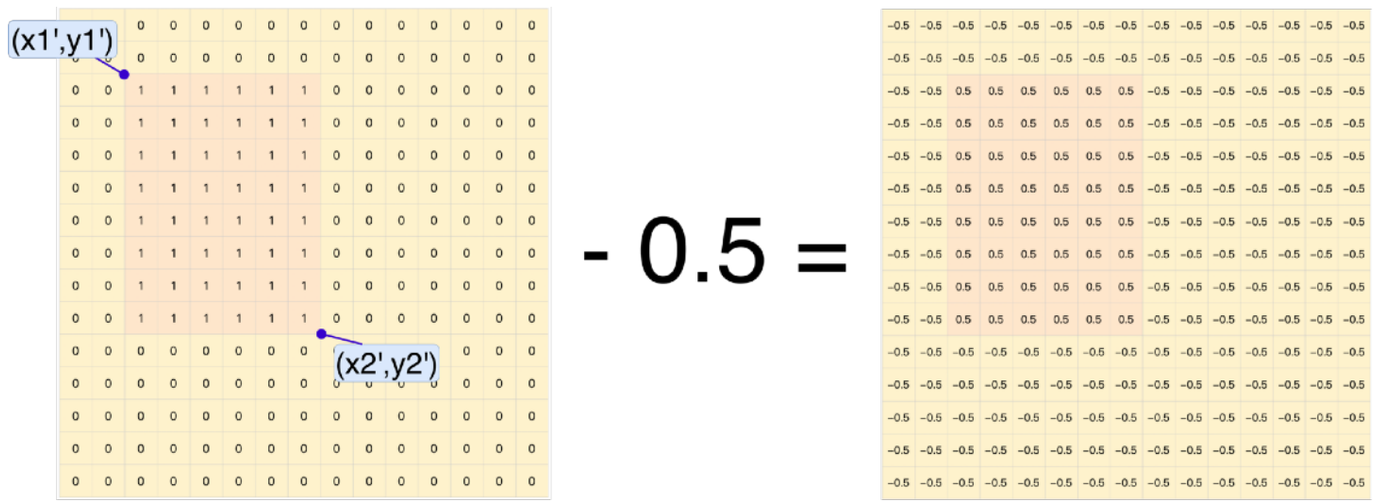
Jing Xu
xujing@tnt.uni-hannover.de

Retina Net

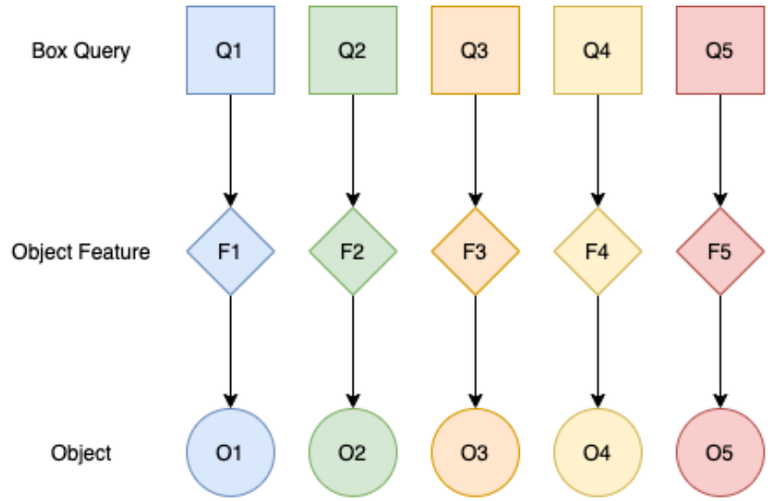
Object Decoder

1.Object Query: box query

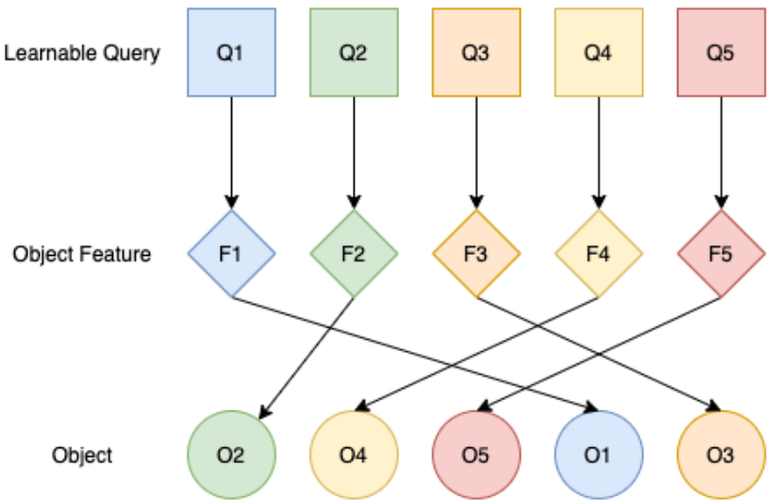




Our query: One-to-one correspondence



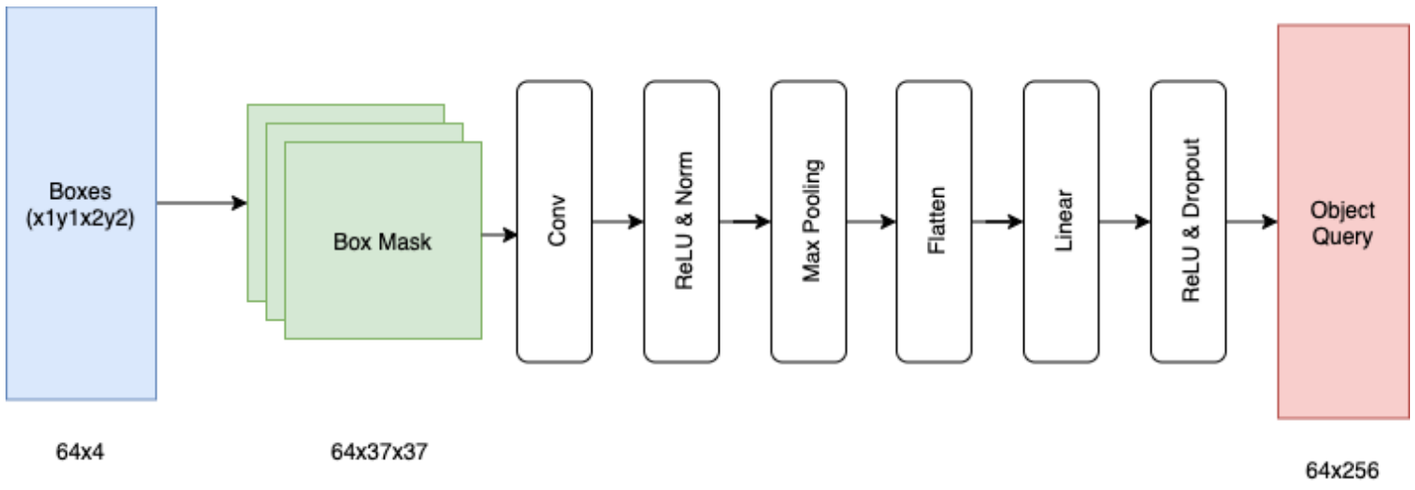
Query in DETR: Need matcher



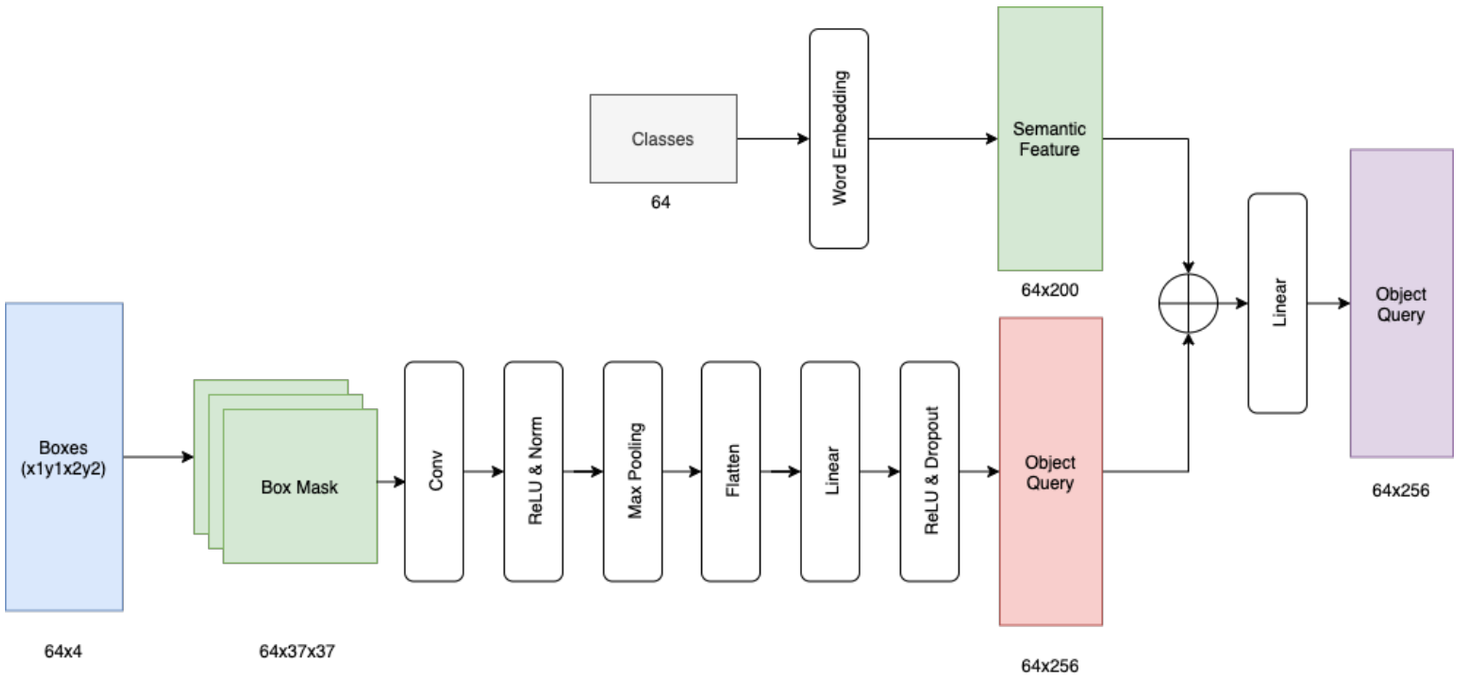
Retina Net

Object Decoder

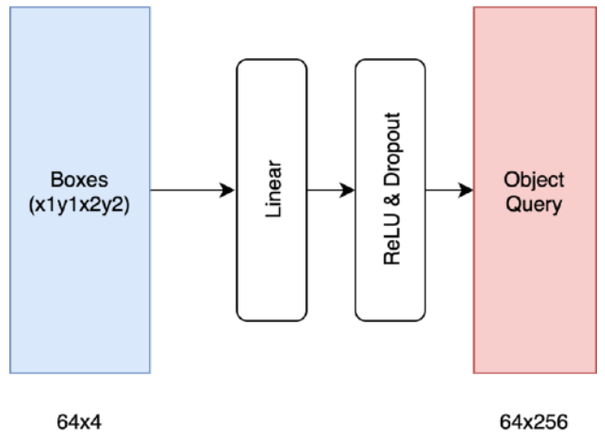
1.Object query



Object query 1



Object query 2: add semantic information



Object query 3: simple design

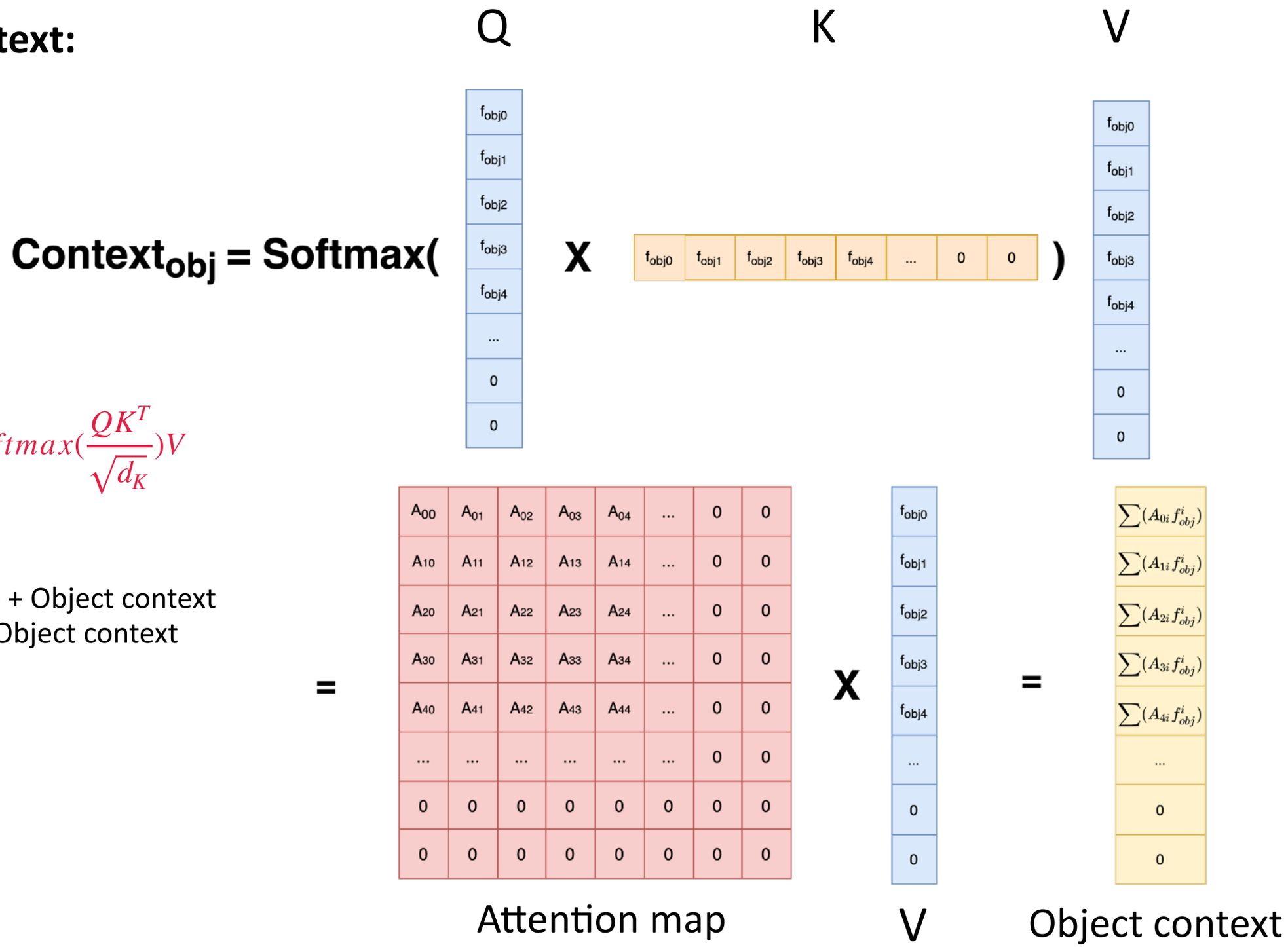
Retina Net

Object Decoder

2.Object Context:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_K}})V$$

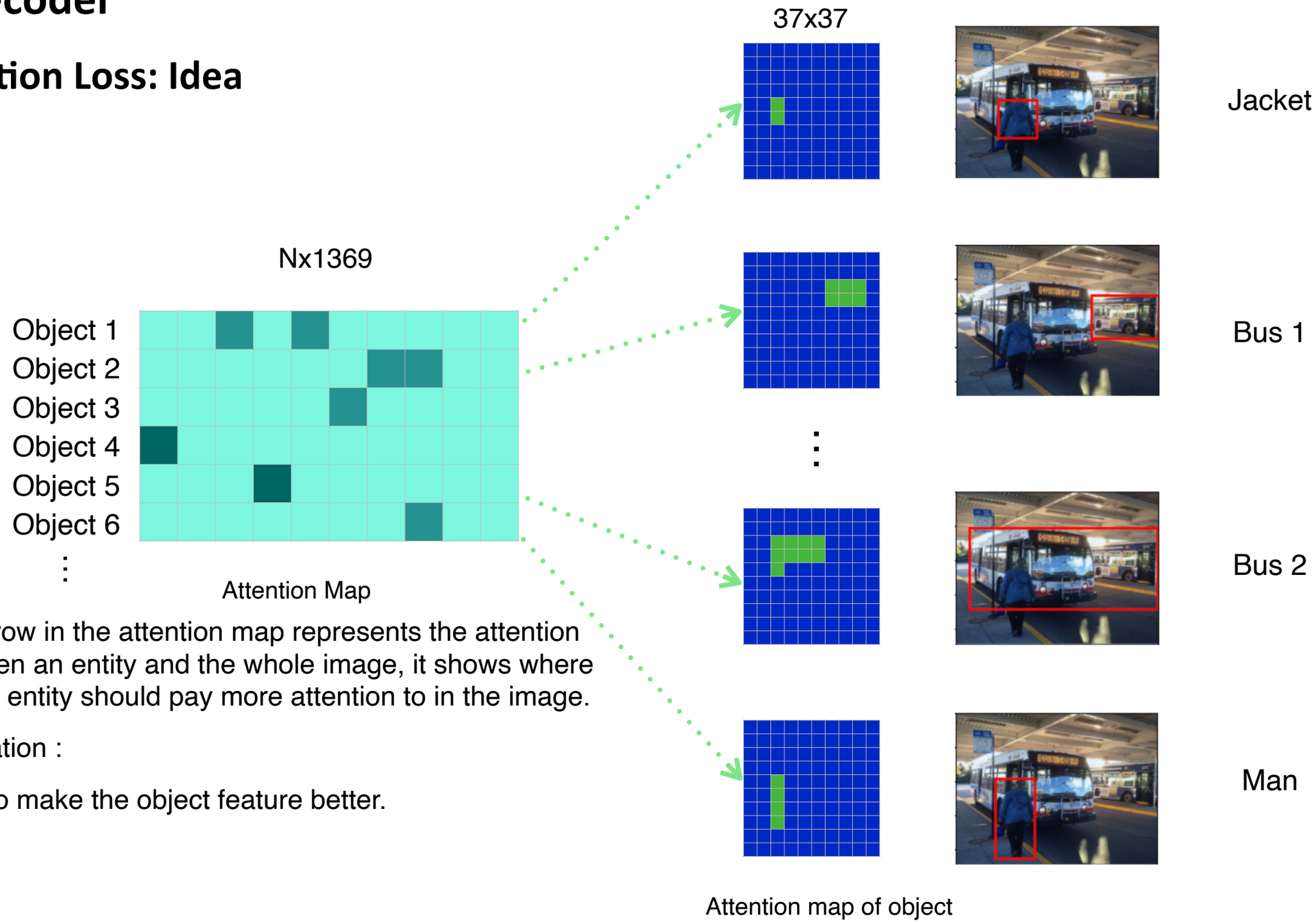
Q(Query): Object feature + Object context
K(Key): Object feature + Object context
V(Value): Object feature



Retina Net

Object Decoder

3.Attention Loss: Idea



Each row in the attention map represents the attention between an entity and the whole image, it shows where for the entity should pay more attention to in the image.

Motivation :

To make the object feature better.

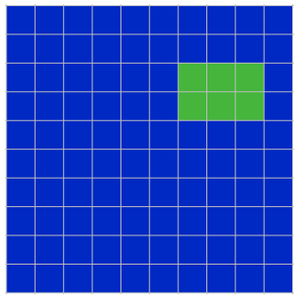
Object Decoder

3. Attention Loss

$$loss_{attention} = max(0, \frac{1}{m}\sum_{m} Att_j^{no\_obj} - \frac{1}{n}\sum_{n} Att_i^{obj} + M)$$

where:

- $Att_j^{no\_obj}$: the j^{th} attention weight of the background.
- Att_i^{obj}: the i^{th} attention weight of ground truth object.
- M: Margin.
- m: the total number of Att^{no_obj}.
- n: the total number of Att^{obj}

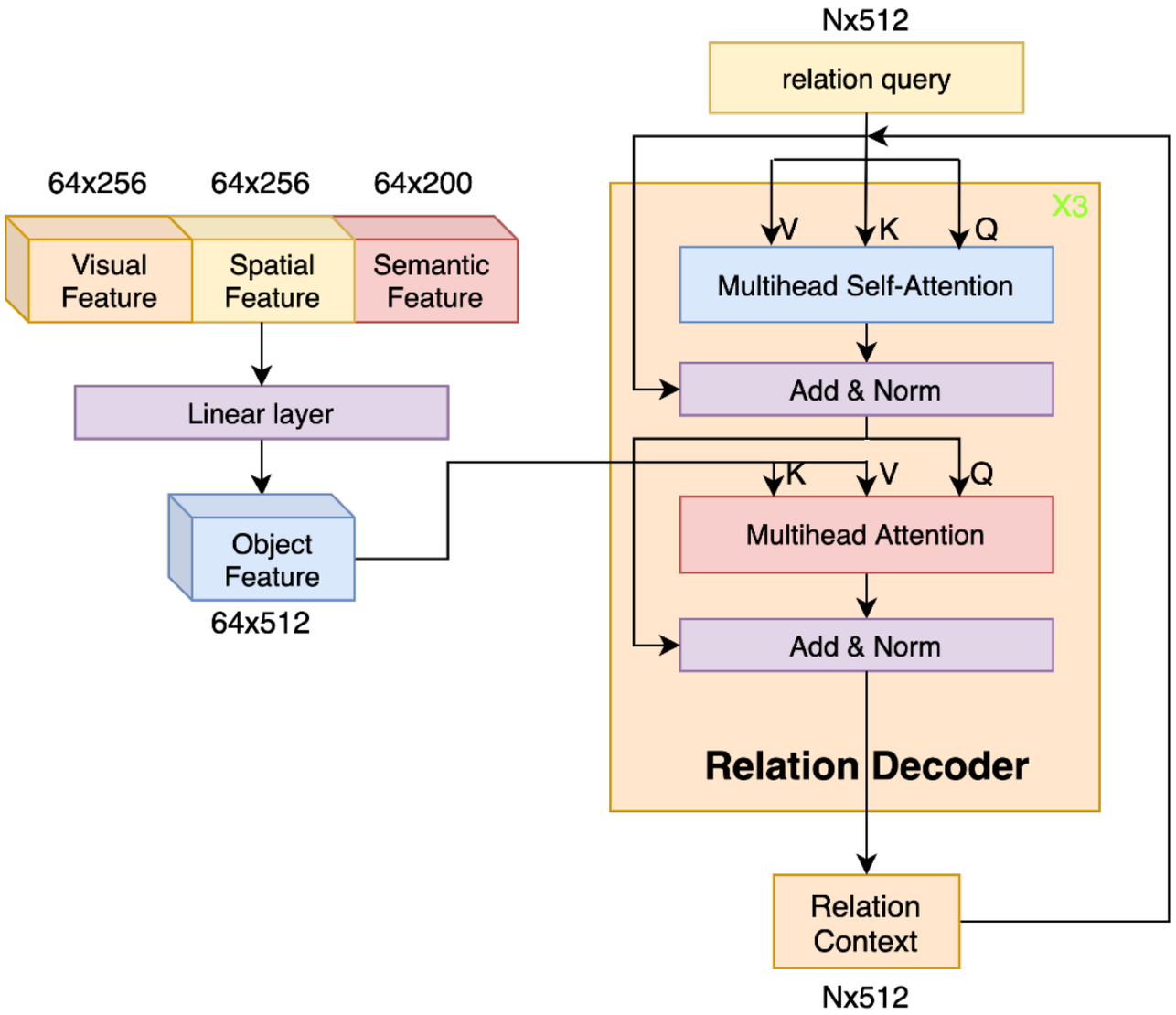


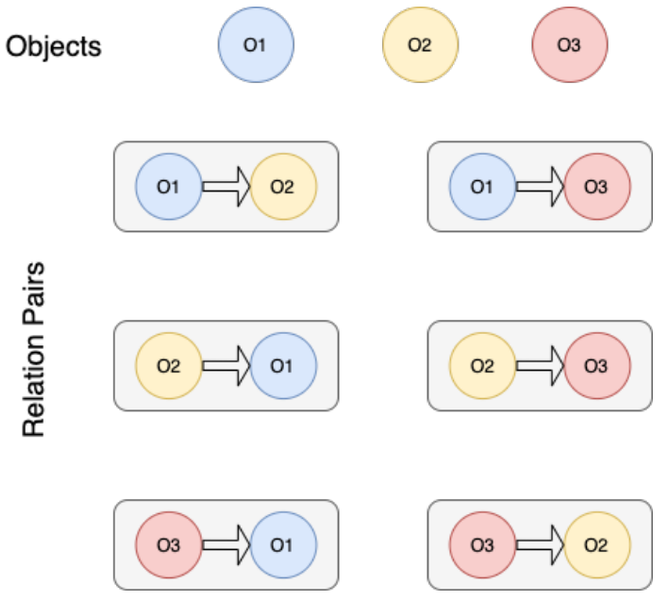
Motivation:

Make Att^{obj} higher, while Att^{no_obj} lower, so that the object feature will pay more attention to object itself.

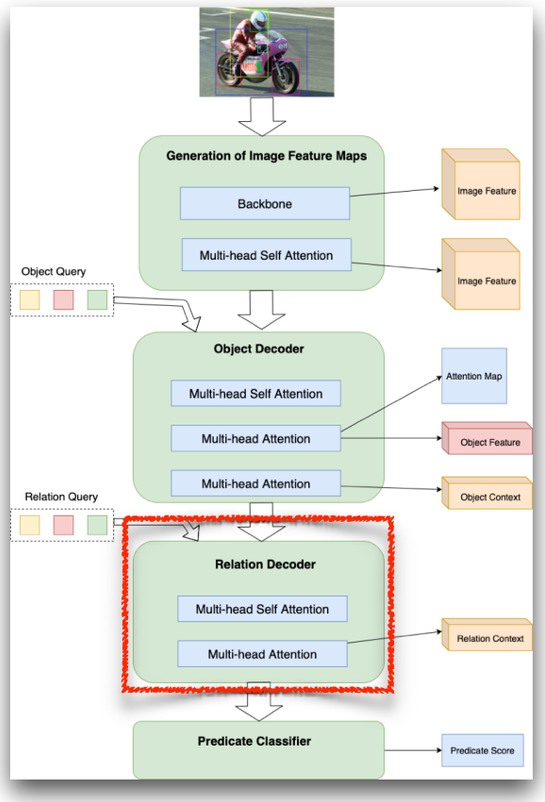
Retina Net

Relation Decoder





K Objects has K(K-1) relation pairs. N=K(K-1)

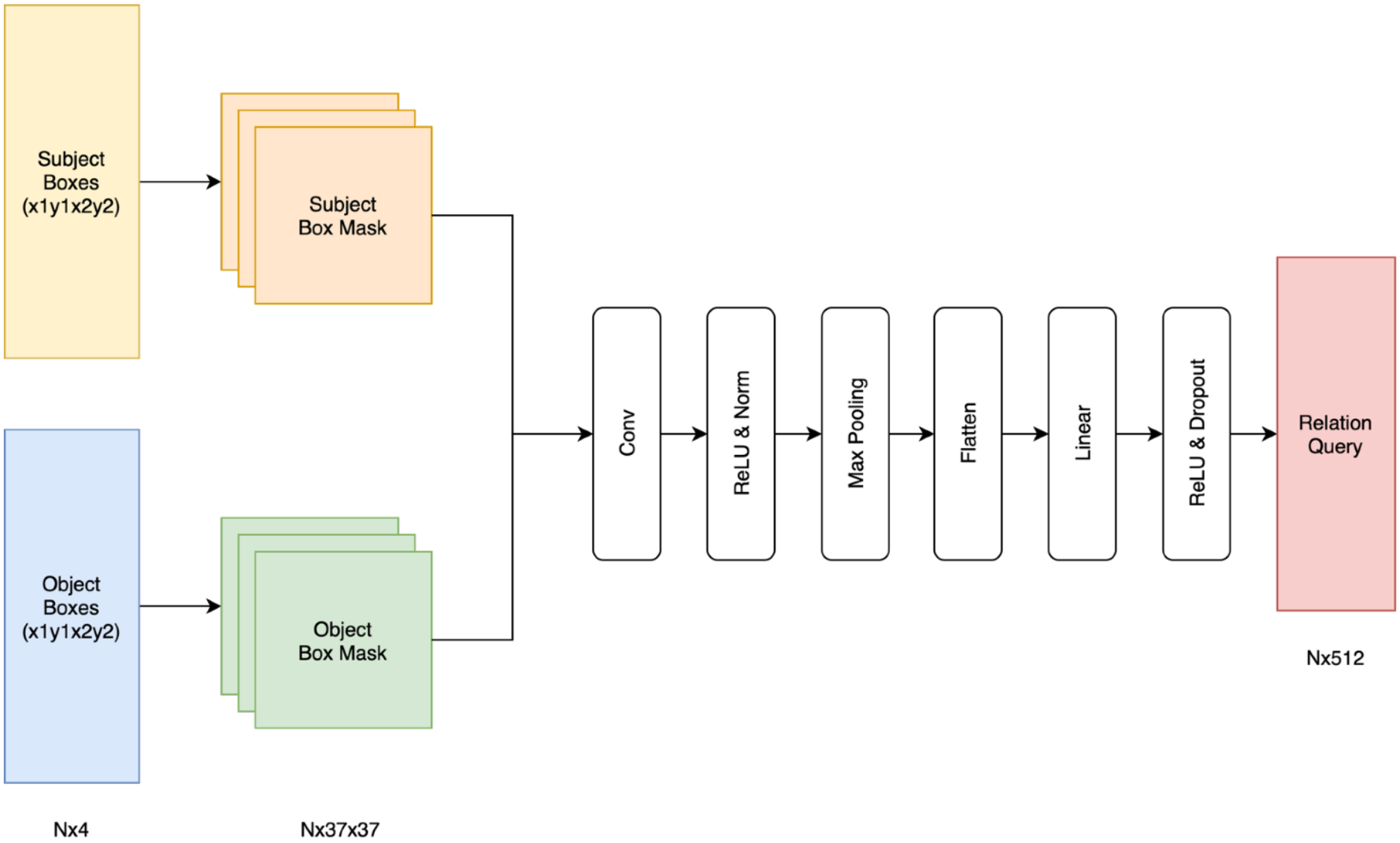


Visual Feature: obtained from object decoder.
Spatial Feature: extracted from bounding boxes.
Semantic Feature: extracted by encoding the object classes through GloVe.

Retina Net

Relation Decoder

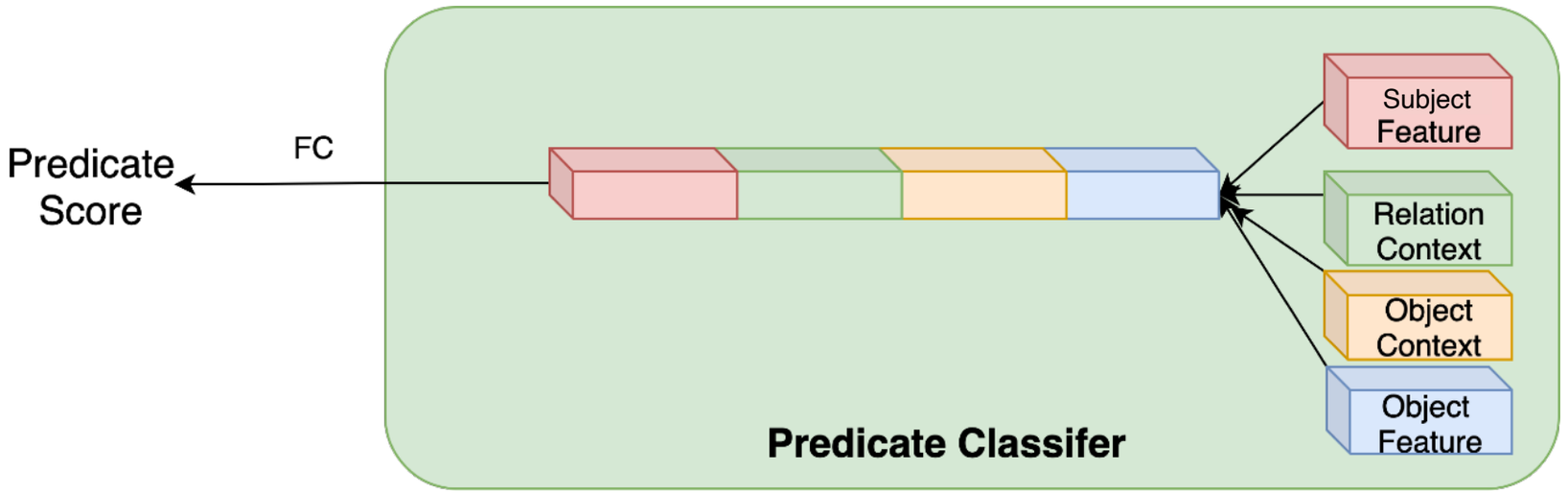
Relation Query

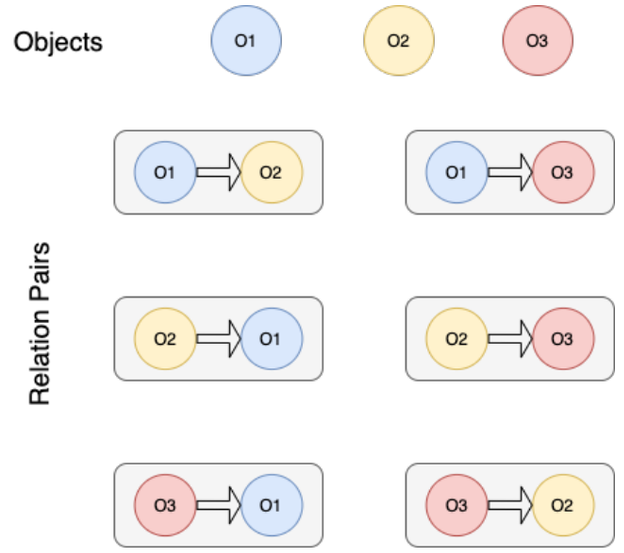


The generation of the relation query is similar to that of the object query, the different is that the input of conv layer is two-channels box mask.

Retina Net

Predicate Classifier



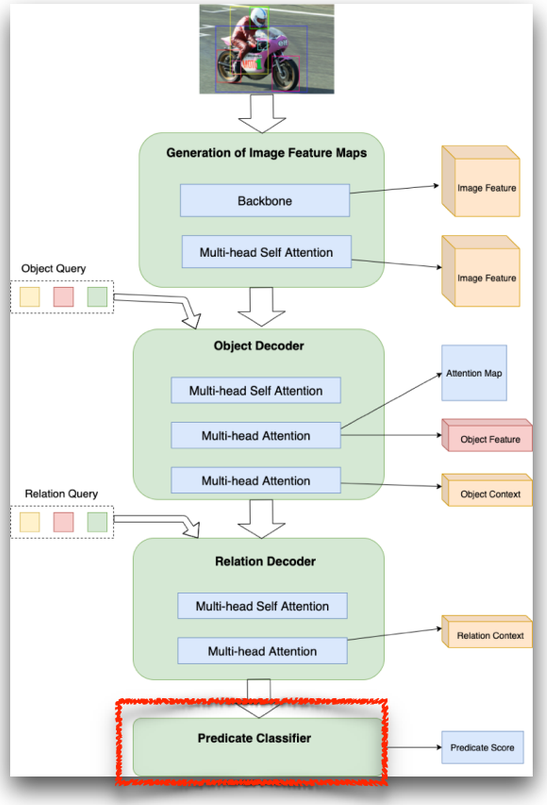


$$score_{pred}(i,j) = Linear(concat(feat^{i}_{sbj}, context^{ij}_{rel}, context^{i}_{obj}, feat^{j}_{obj})$$

Where:

i is the index of subject,

j is the index of object.



Outline

Introduction

Our Method

Dataset and Evaluation Metrics

- Visual Genome

Experimental Results

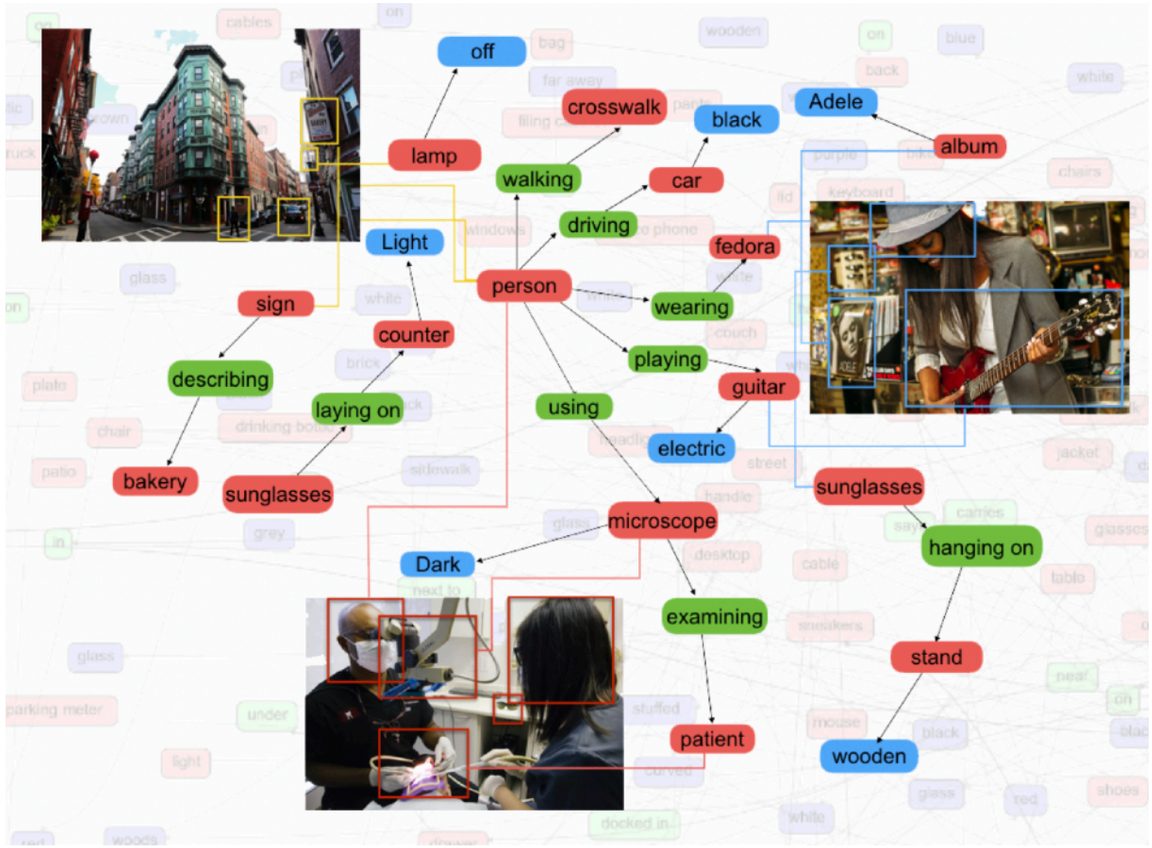
Conclusions

VG Dataset

Visual Geneme

In visual relationship detection / scene graph generation, most of related works adopt Visual Genome as dataset.

- 108,077 Images
- 3.8 Million Object Instances
- 2.3 Million Relationships



Our dataset

- 62723 images for train dataset
- 26446 images for val dataset

Evaluation Metrics

- Recall @ K
- PredCLS, SGCLS, SGDET

Outline

Introduction

Our Method

Dataset and Evaluation Metrics

Experimental Results

- Object decoder
- Relation decoder
- Setting of transformer
- Qualitative Results

Conclusions

Object Decoder

1.Ablation Study for Object query

	Object query 1	Object query 2	Object query 3
IoU	**0.759**	0.750	0.681
GIoU	**0.744**	0.733	0.669
	Object feature 1	**Object feature 2**	**Object feature 3**
IoU	**0.634**	0.631	0.580
GIoU	0.617	**0.620**	0.568

Table 1: The regression result of the object queries and features.

	Object query 1	Object query 2	Object query 3
Recall@50	62.9	**63.2**	63.0
Recall@100	65.0	**65.3**	65.1

Table 2: The comparison of the object queries in PredCLS.

1. The object query and object feature we designed correspond to each other.
2. The performance of the three object queries are similar.

Object Decoder

2.Ablation Study for Object context

	Recall@20	Recall@50	Recall@100
Without object context	51.7	59.6	62.3
With object context	**55.4**	**63.4**	**64.4**

The effect of object context in PredCLS

The object context has great benefits for the predicate prediction.

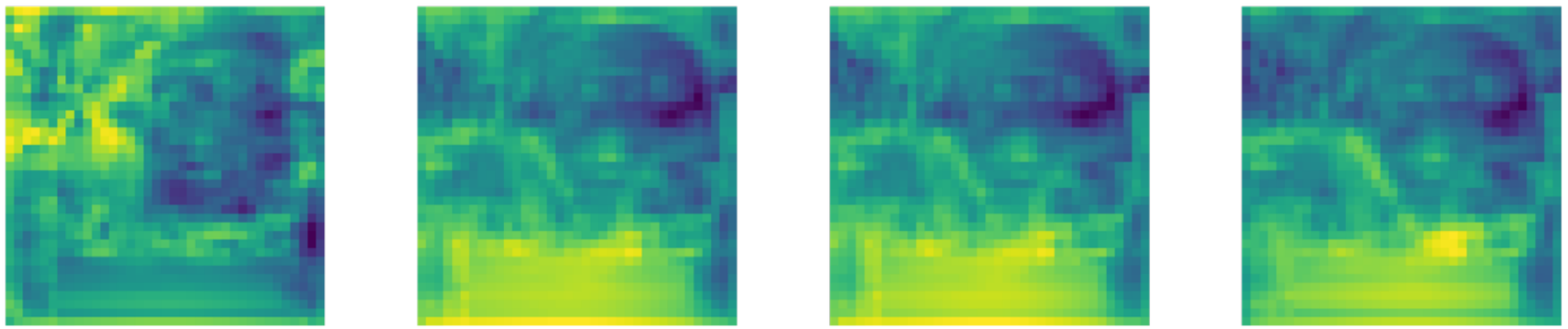
Retina Net

Object Decoder

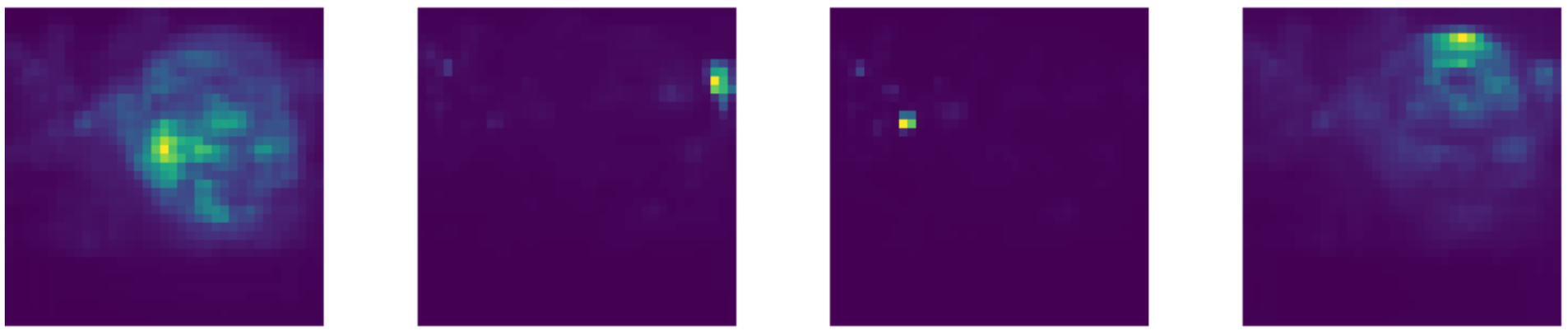
3.Visualised results of Attention Loss

Objects

Attention map without our loss



Attention map with our loss



Object Decoder

3.Ablation study for Attention Loss

	PredCLS			SGCLS		
	Recall@20	**Recall@50**	**Recall@10**	**Recall@20**	**Recall@50**	**Recall@100**
Without our attention loss	53.9	61.8	64.0	25.7	30.7	31.4
With our attention loss	**55.4**	**63.3**	**65.2**	**28.4**	**32.7**	**33.6**

The result of our attention loss in PredCLS and SGCLS

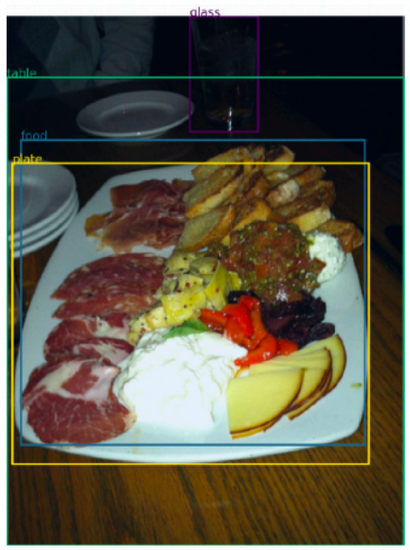
1. Our attention loss can effectively change the attention map so that it can be well visualized and can show the position and shape of each object.

2. The model with our attention loss can achieve better performance.

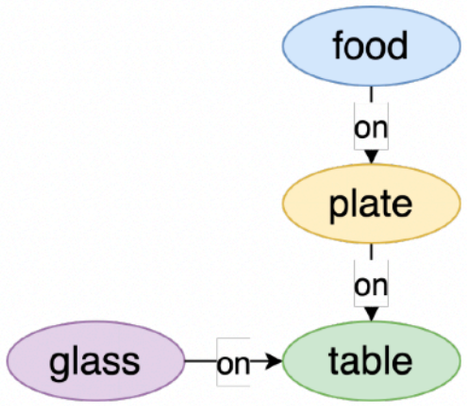
Retina Net

Relation Decoder

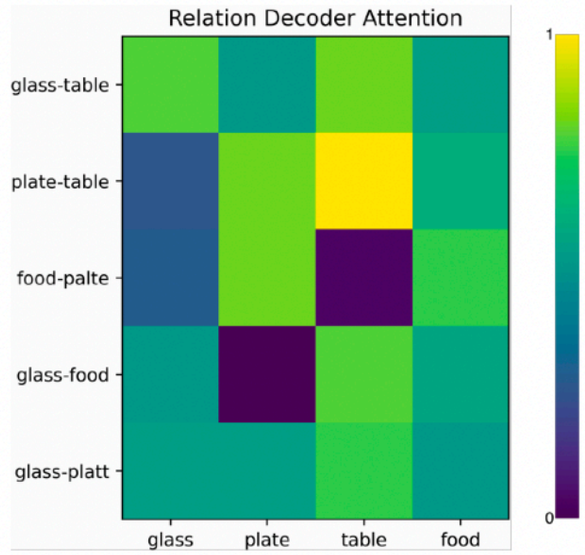
Visualised results of Relation Context



(a) Scene graph

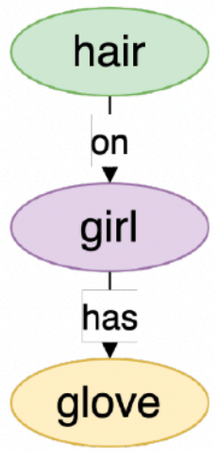


(b) Ground truth relationships

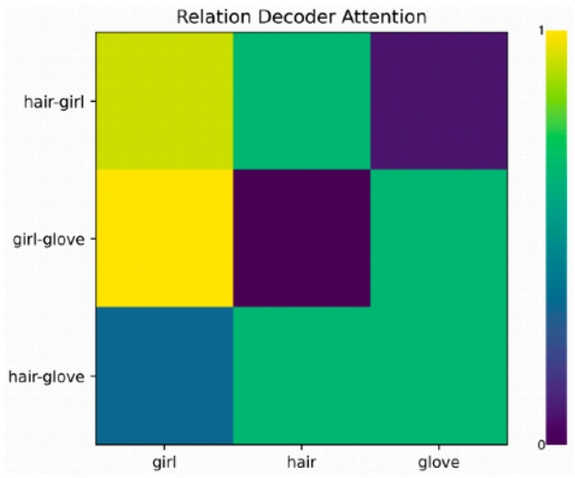


(c) Attention map

(d) Scene graph



(e) Ground truth relationships



(f) Attention map

Ablation study for Relation Decoder

	Recall@20	Recall@50	Recall@100
Without relation decoder	51.1	58.7	61.1
With relation decoder	**55.7**	**63.4**	**65.4**

Table 1: The effect of our relation decoder in PredCLS

	Recall@20	Recall@50	Recall@100
Only visual	53.56	61.10	62.21
Visual + spatial	54.44	61.96	63.82
Visual + semantic	54.76	62.73	64.35
Visual + spatial + semantic	**55.67**	**63.40**	**65.41**

Table 2: The effect of the expression of object feature in the relation decoder in PredCLS

1. The relation decoder is very helpful to solve the visual relation detection problem.

2. The expression of object feature can affect our work. When the object feature has visual, spatial and semantic information, it has best result.

Setting of Transformer

Layer	Head	Recall@20	Recall@50	Recall@100
1	4	55.34	63.10	65.11
2	**4**	55.50	63.19	65.21
3	**4**	55.51	63.25	65.25
4	**4**	55.54	63.35	65.36
2	**8**	55.63	63.38	65.38
3	**8**	**55.67**	**63.40**	**65.41**

Table 1: The setting of transformer

	PredCLS		SGCLS	
	Recall@50	**Recall@100**	**Recall@50**	**Recall@100**
With encoder	62.1	63.2	**33.7**	**34.6**
Without encoder	**63.4**	**64.4**	32.8	33.4

Table 2: The result of encoder in PredCLS and SGCLS

1. The attention layers and attention heads don’t have a great influence on predicate prediction.

2. The PredCLS is more suitable for without encoder, and the SGCLS is more suitable for with encoder.

Retina Net

Results Comparison

Models	PredCLS			SGCLS			SGDET		
	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
MESSAGE PASSING	-	44.8	53.1	-	21.7	24.4	-	3.4	4.2
ASSOC EMBED	47.9	54.1	55.4	18.2	21.8	22.6	6.5	8.1	8.2
MSDN	-	42.3	48.2	-	20.9	24.0	-	11.7	14.0
FREQ	53.6	60.6	62.2	29.3	32.3	32.9	20.1	26.2	30.1
Motif NET	**58.5**	**65.2**	**67.1**	**32.9**	**35.8**	**36.5**	**21.4**	27.2	30.3
KERN	-	65.8	67.6	-	36.7	37.4	-	27.1	29.8
RelDN	-	-	-	-	-	-	21.1	**28.3**	**32.7**
Ours	55.4	63.4	65.4	29.4	33.7	34.6	20.2	26.3	29.6

Comparison with some advanced realted works.

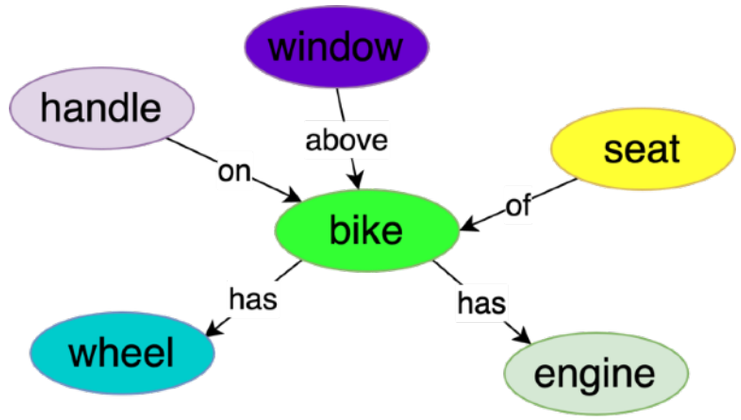
Our Retina Net can complete the visual relation detection problem well.

Retina Net

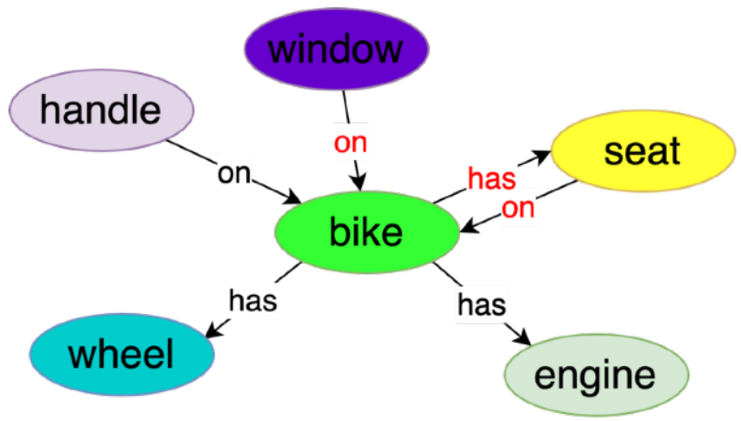
Qualitative Results

Predicate classification

Scenes graph with bounding boxes and class labels



The ground truth relationships



The result of Recall@50

The recall@50: 60.0%.

The wrong relationships: < window on bike > , < seat on bike > and < bike has seat >.

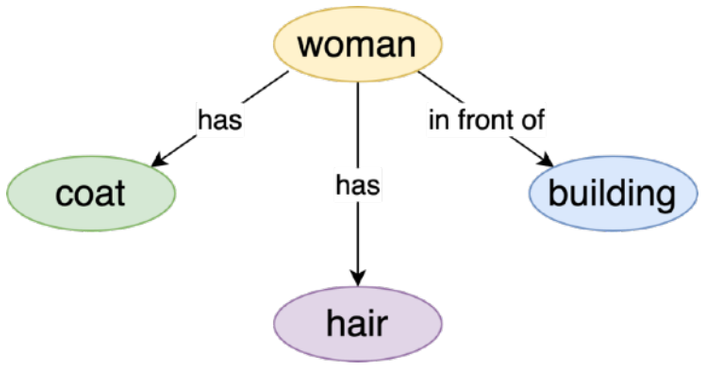
The possible reasons: predicates such as *'on'* and *'has'* are the majority in our data set, so our model predicts them into these.

Retina Net

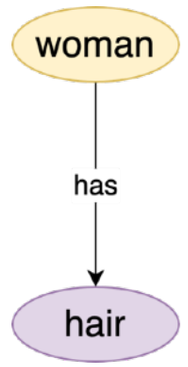
Qualitative Results

Scene graph classification

Scenes graph with bounding boxes and class labels



The ground truth relationships



The result of Recall@50

The recall@50: 33.3%.

The possible reason: the model did not predict *'building'* and *'coat'*. The score of *'building'* is only 0.72% and the score of *'coat'* is only 3.4%.

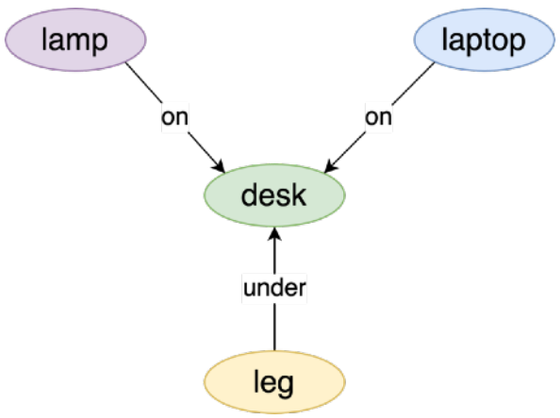
Retina Net

Qualitative Results

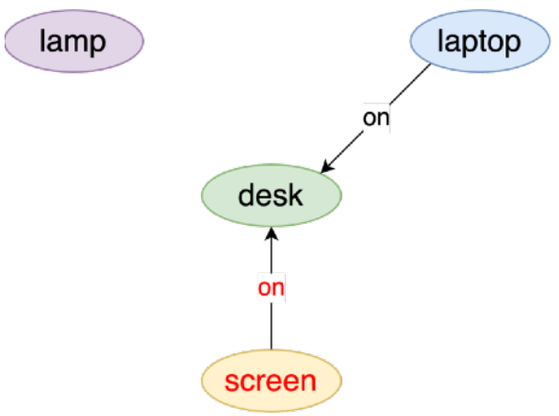
Scene graph detection

Scenes graph with bounding boxes and class labels

Scenes graph with dectected boxes and class labels



The ground truth relationships



The result of Recall@50

The recall@50: 33.3%.

The detected box of *'lamp'* is much smaller than the ground truth, and the IoU between them is 0.23.

The relationship <screen on desk> is correct but not in the ground truth relationships.

Outline

Introduction

Our Method

Dataset and Evaluation Metrics

Experimental Results

Conclusions

Conclusions

Conclusions:

- Our model Retina Net based on transformer can solve visual relation detection task well.
- Our object query is very suitable for predicate classification and scene graph classification.
- The attention loss can complete the task very well, which makes our object feature more recognisable and the object can be visualised through the corresponding attention map.
- We propose the global object context and relation context , which obviously improved the final results.

Thank you!

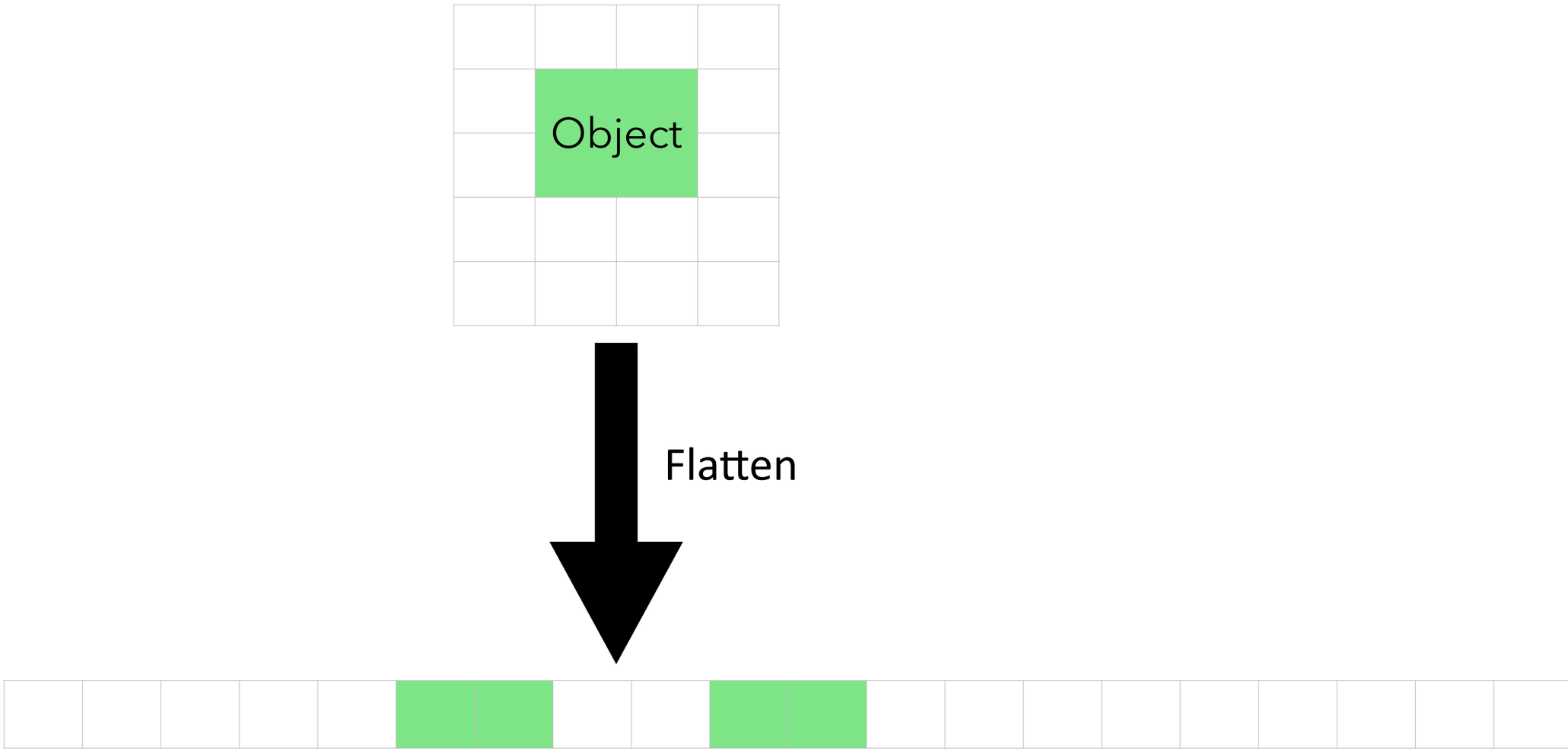
$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_K}})V$$

The sum of attention weights of every row is 1, because of softmax.

The average value of all attention weighs: 1/N

Margin: 1/4N

$$loss_{attention} = max(0, \frac{1}{m}\sum_{m} Att_j^{no\_rel} - \frac{1}{n}\sum_{n} Att_i^{rel} + M)$$

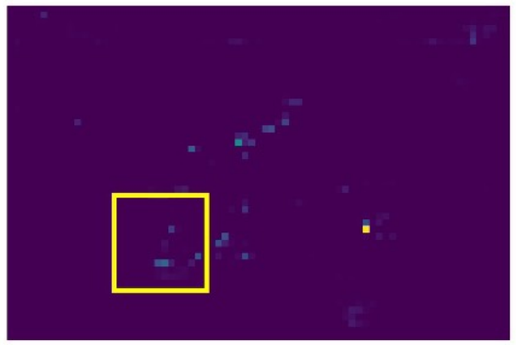

Object
Flatten

Pixel-based Attention

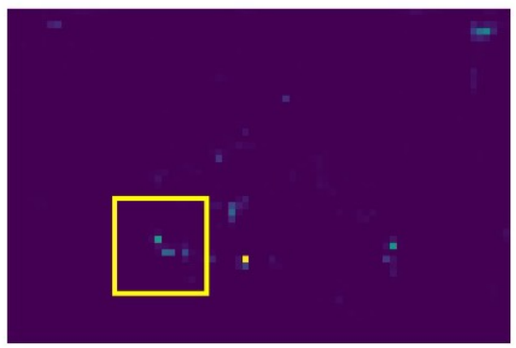
Visualised results

Image: motorcycle has wheel

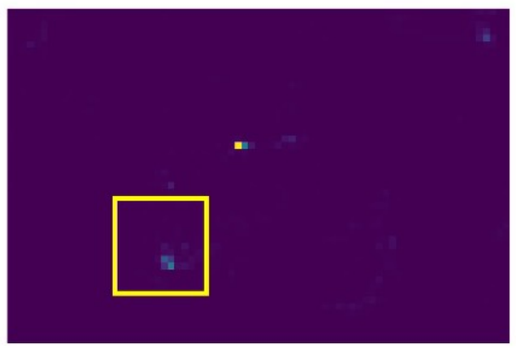
Pixel attention map (1)



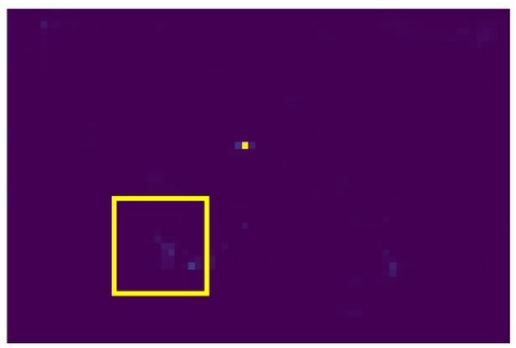
Pixel attention map (2)



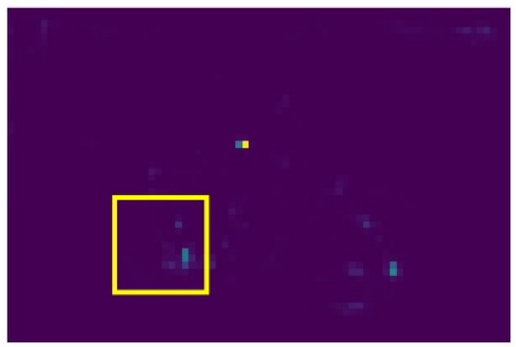
Pixel attention map (3)



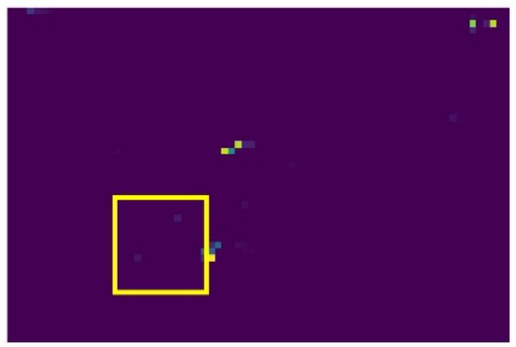
Pixel attention map (4)



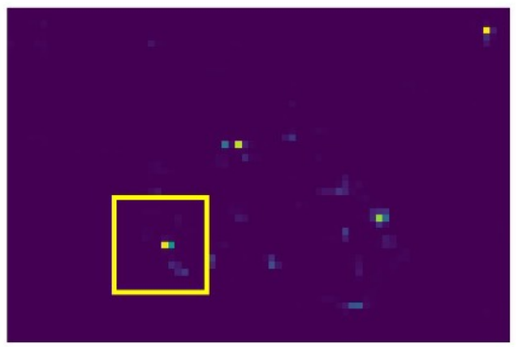
Pixel attention map (5)



Pixel attention map (6)



Pixel attention map (7)



Pixel attention map (8)

Relation pair: <motorcycle has wheel>
Red box: subject <motorcycle>
Yellow box: object <wheel>

Result:

The subject <motorcycle> pay attention to the object <wheel>, it also pay attention to the other part of the subject.

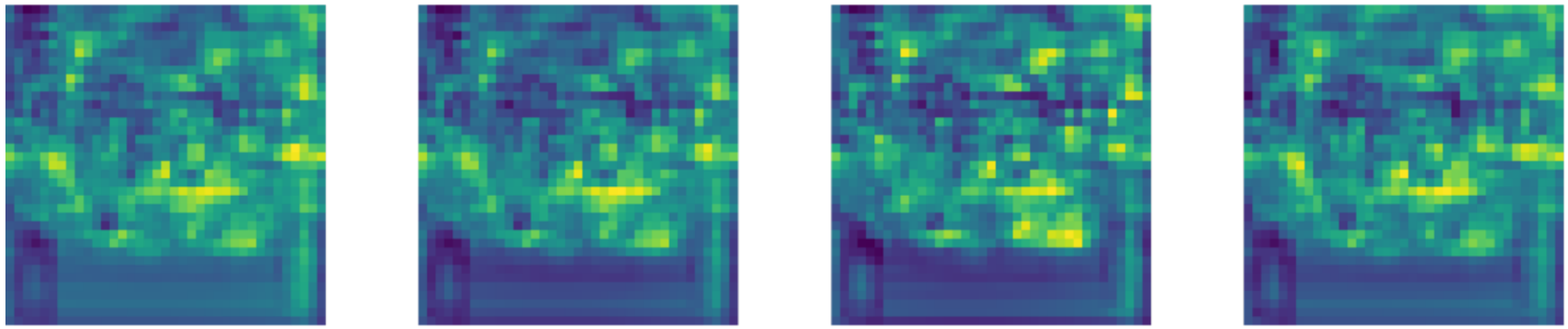
Retina Net

Attention Loss

Visualised results

Objects

Attention map without our loss



Attention map with our loss