# Visual Tracking and Retrieval by Natural Language Descriptions

Qi Feng

Boston University - Image and Video Computing
fung@bu.edu

# A conventional visual tracker

# Visual Tracker

- Target Initialization
  - Bounding box for the first frame.
- Appearance Modeling
  - Conventional & deep visual features.
- Motion Estimation
- Object Localization

*Li, Xi, et al. "A survey of appearance models in visual object tracking." ACM TIST 4.4 (2013): 58.*

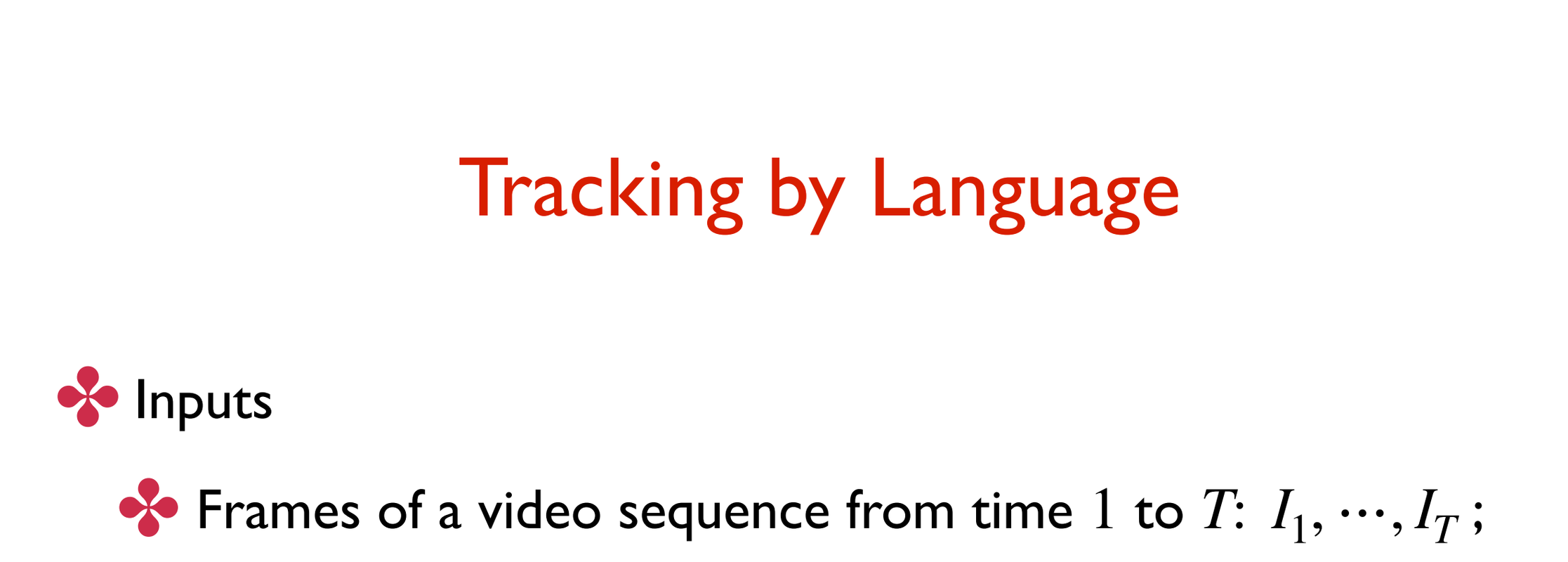
# NL Tracker

- Target Initialization
  - Bounding box for the first frame;
  - NL description $Q$.
- Appearance Modeling
  - Conventional & deep visual features;
  - Language features as template.
- Motion Estimation
- Object Localization

*Li, Zhenyang, et al. "Tracking by natural language specification." CVPR. 2017.*
*Feng, Qi, et al. "Real-time visual object tracking with natural language description." WACV. 2020.*

# Tracking by Language

- Inputs
  - Frames of a video sequence from time 1 to $T$: $I_1, \cdots, I_T$ ;
  - NL Description $Q$ of the target.

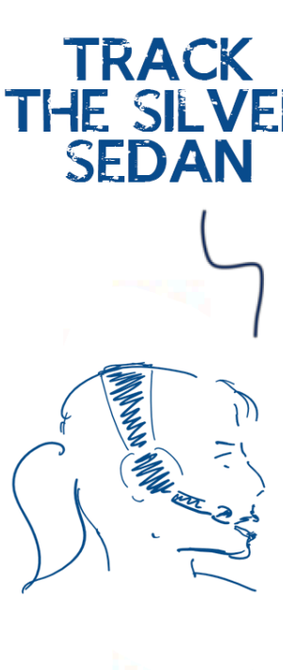


# Tracking by NL
# The Problem Definition

- Target Initialization
  - Automatic: from $Q$;
  - Additional manual initialization: Bounding Boxes $X^*$, *etc.*


# Tracking by NL
# The Problem Definition

- Outputs
  - Spatial and Temporal localization of the target;
  - *i*.e. a sequence of bounding boxes, $\hat{X}_{t_1}, \cdots, \hat{X}_{t_2}$.



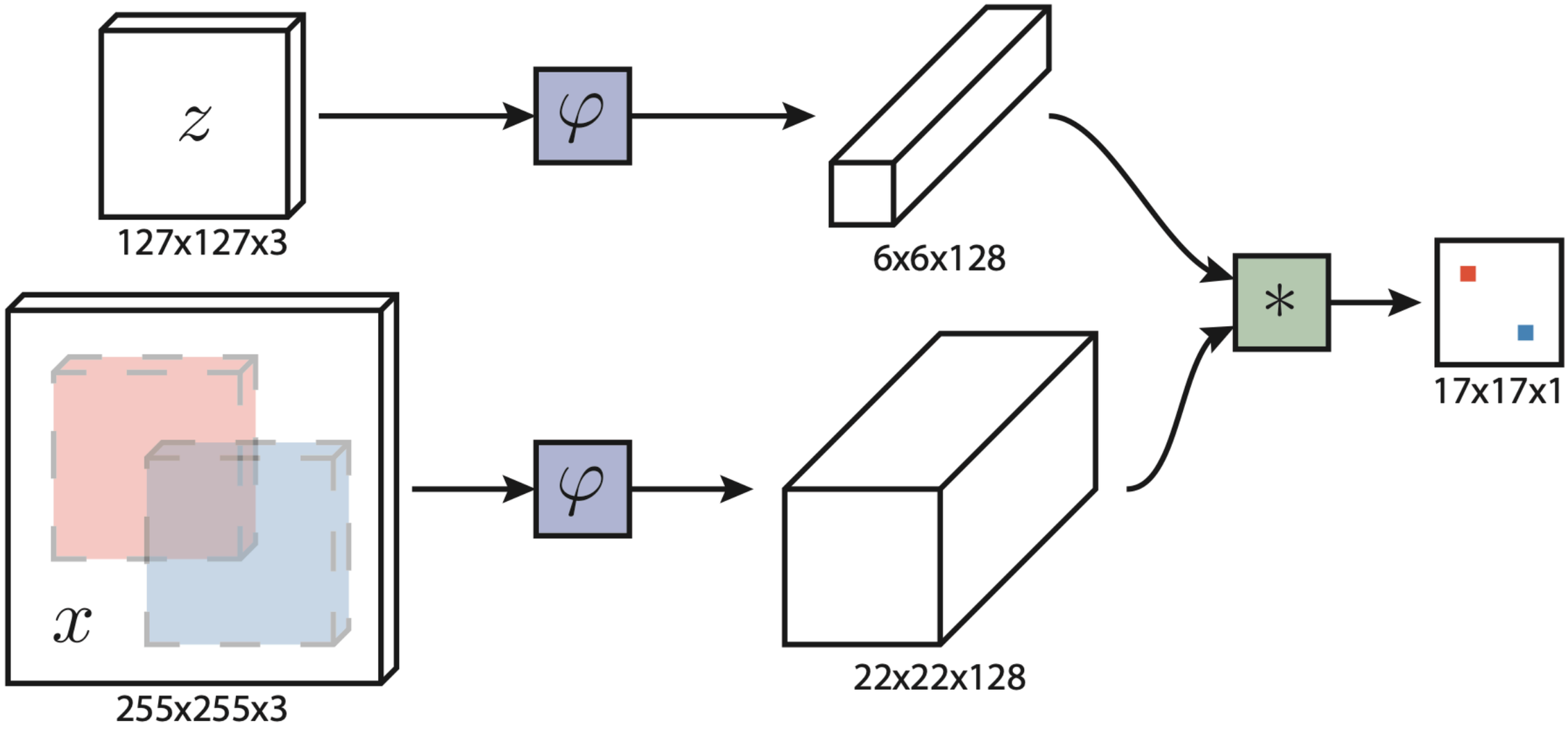

# Ultimate Goal:

# An AI system that detects and tracks targets based on given language instructions

# This Lecture

- Tracking By NL - single object - siamese approach
- CityFlow-NL Dataset
- Tracked-vehicle retrieval by NL

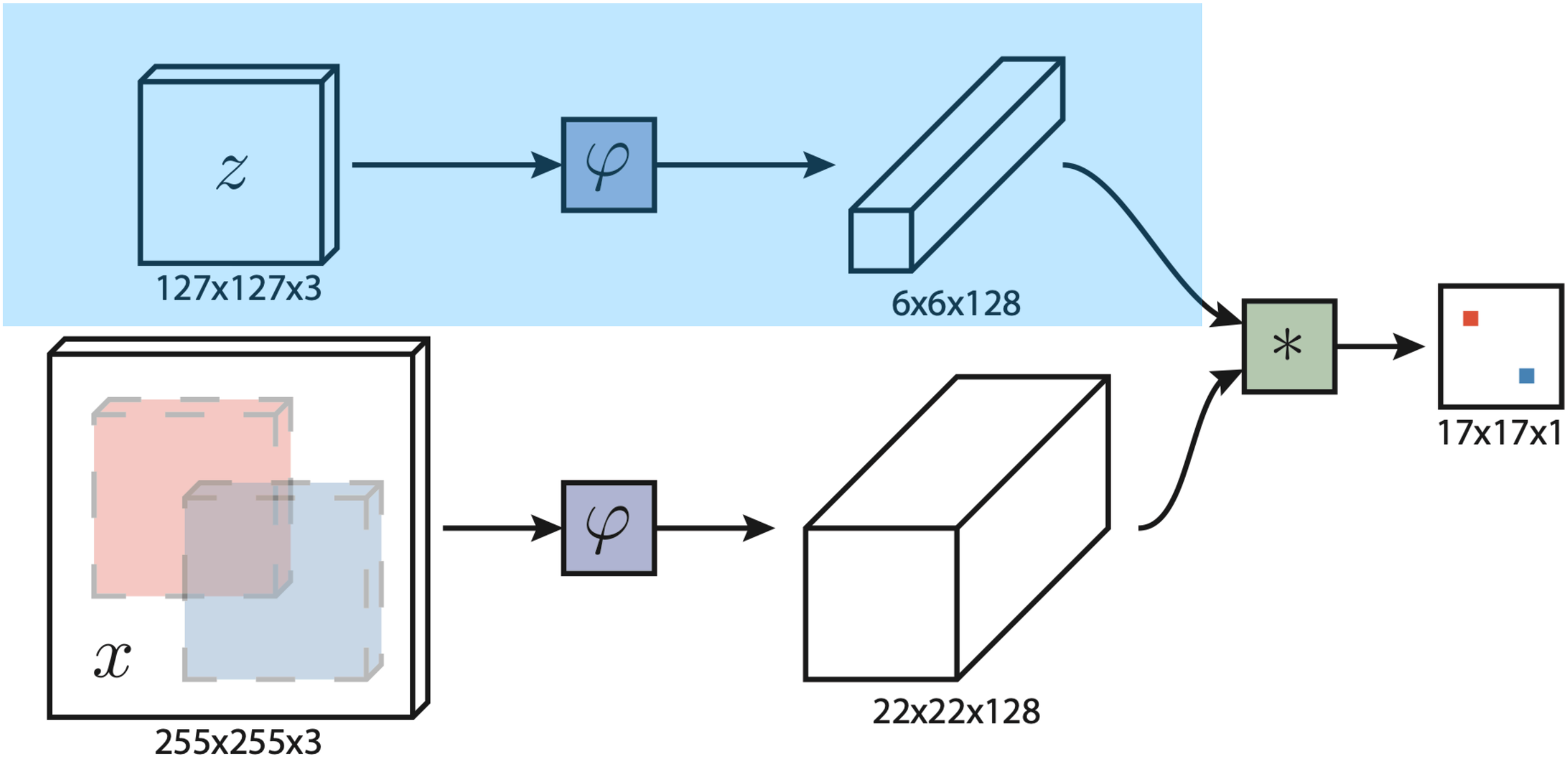
# From last time: SiamFC




*Bertinetto, Luca, et al. "Fully-convolutional siamese networks for object tracking." ECCV, 2016.*

**Replace the visual template with Language**

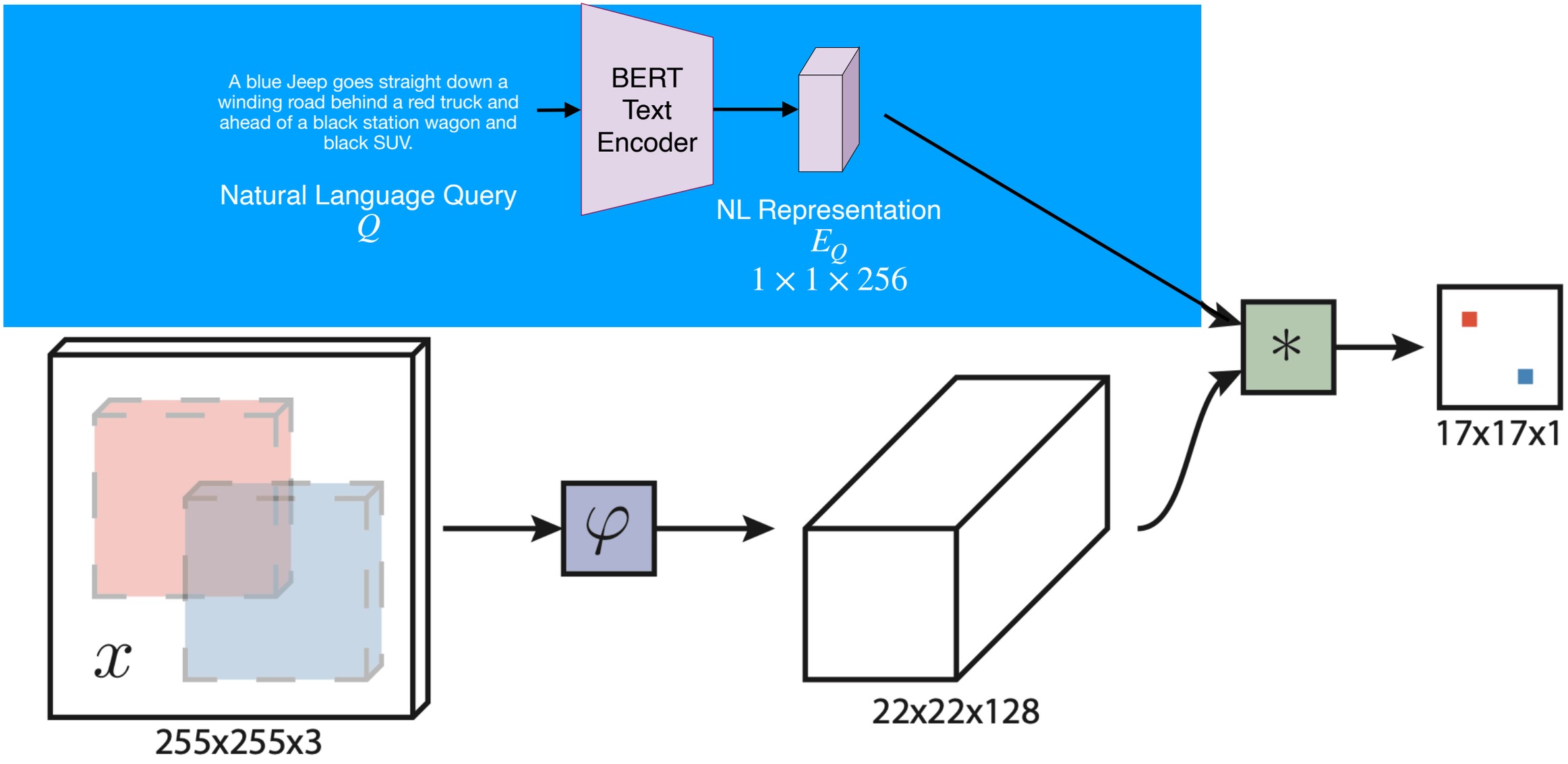
# Our Goal




*Bertinetto, Luca, et al. "Fully-convolutional siamese networks for object tracking." ECCV, 2016.*

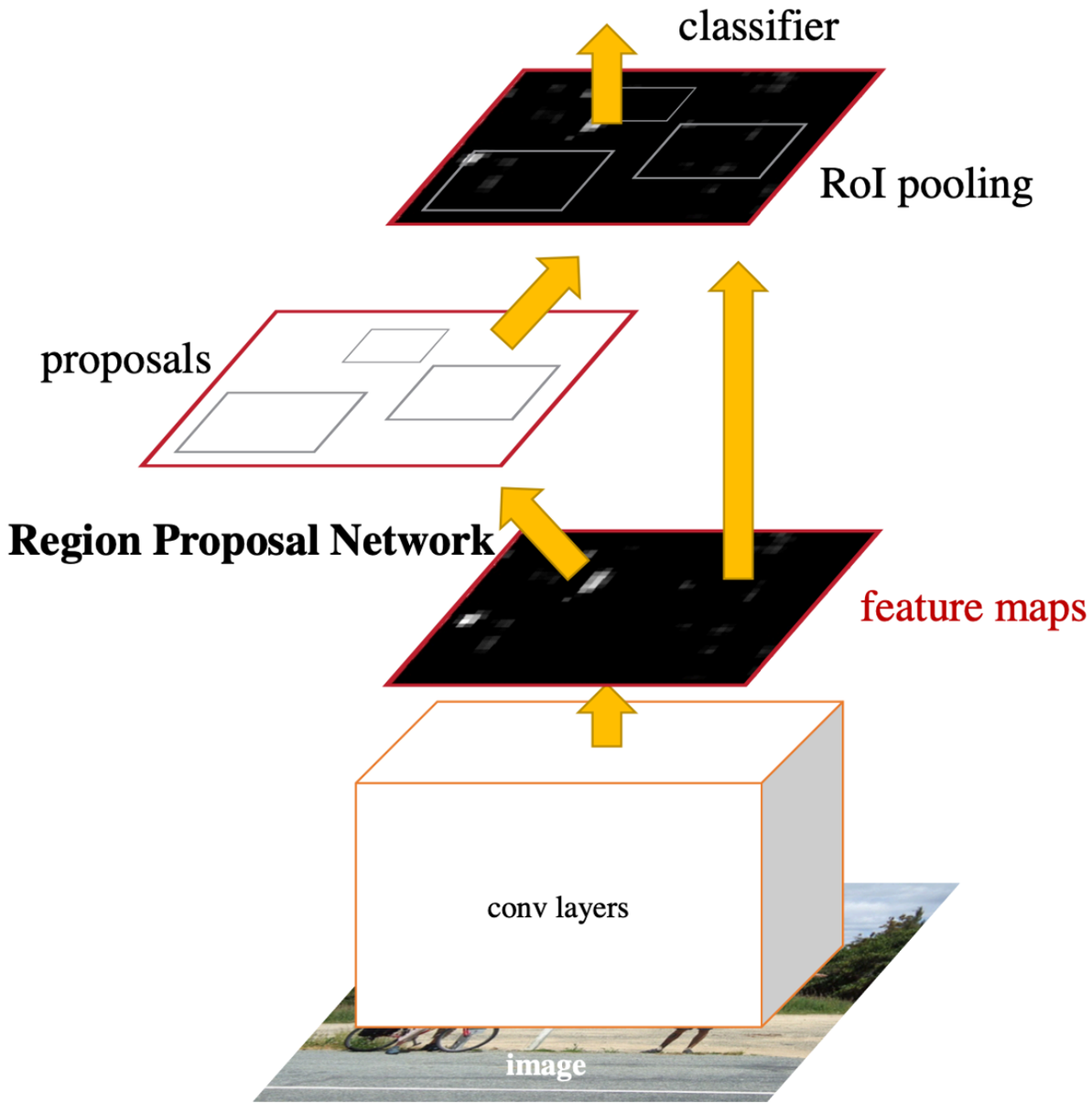
**Replace the visual template with Language**

# Our Goal



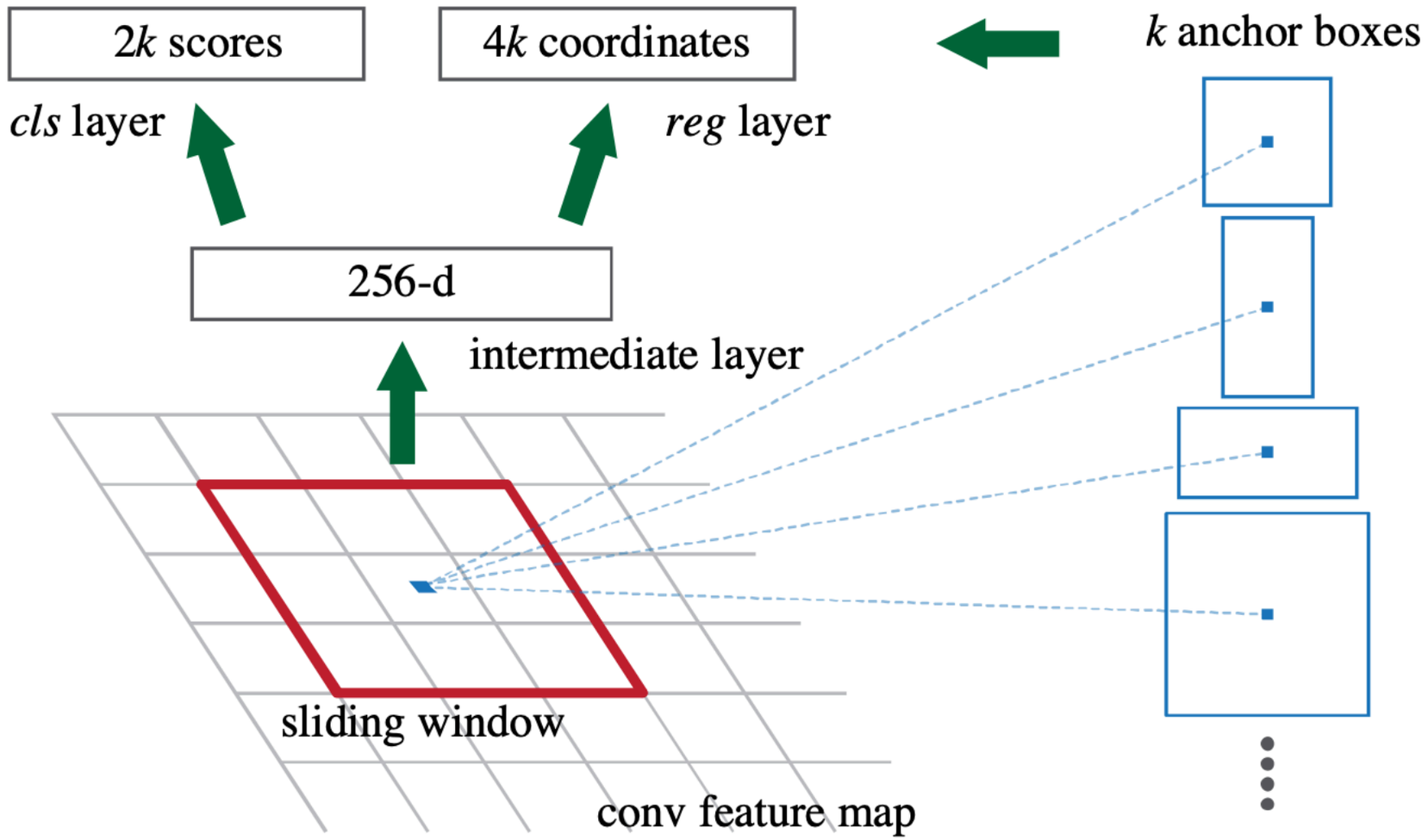

*Bertinetto, Luca, et al. "Fully-convolutional siamese networks for object tracking." ECCV, 2016.*

# Region Proposal Network (RPN)



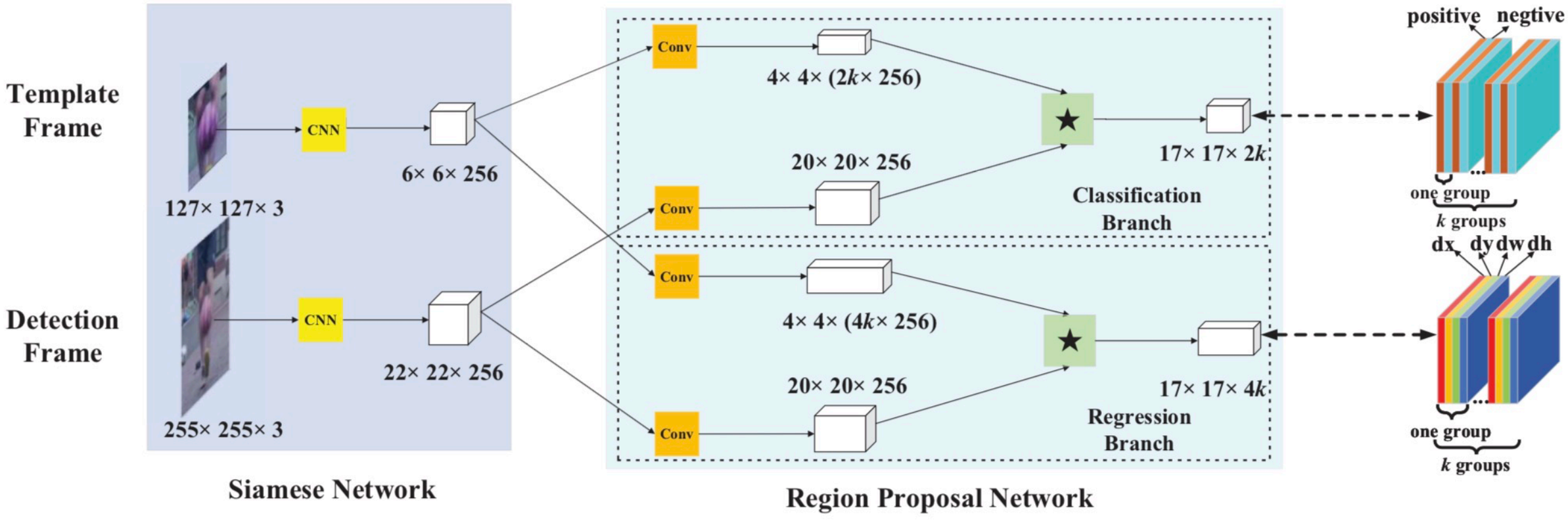

*Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." NeurIPS, 2015.*

# Region Proposal Network (RPN)



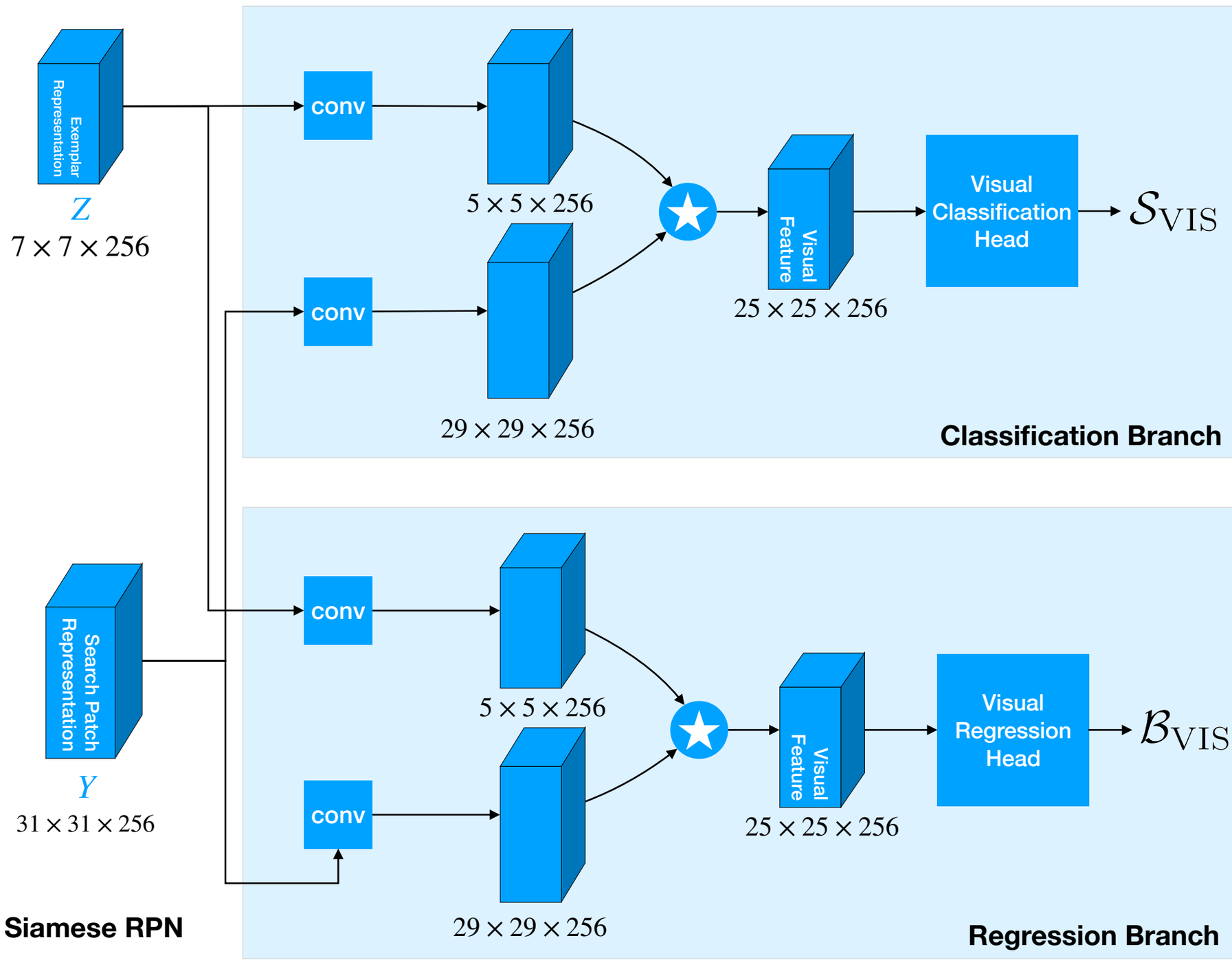

*Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." NeurIPS, 2015.*

# Siamese RPN



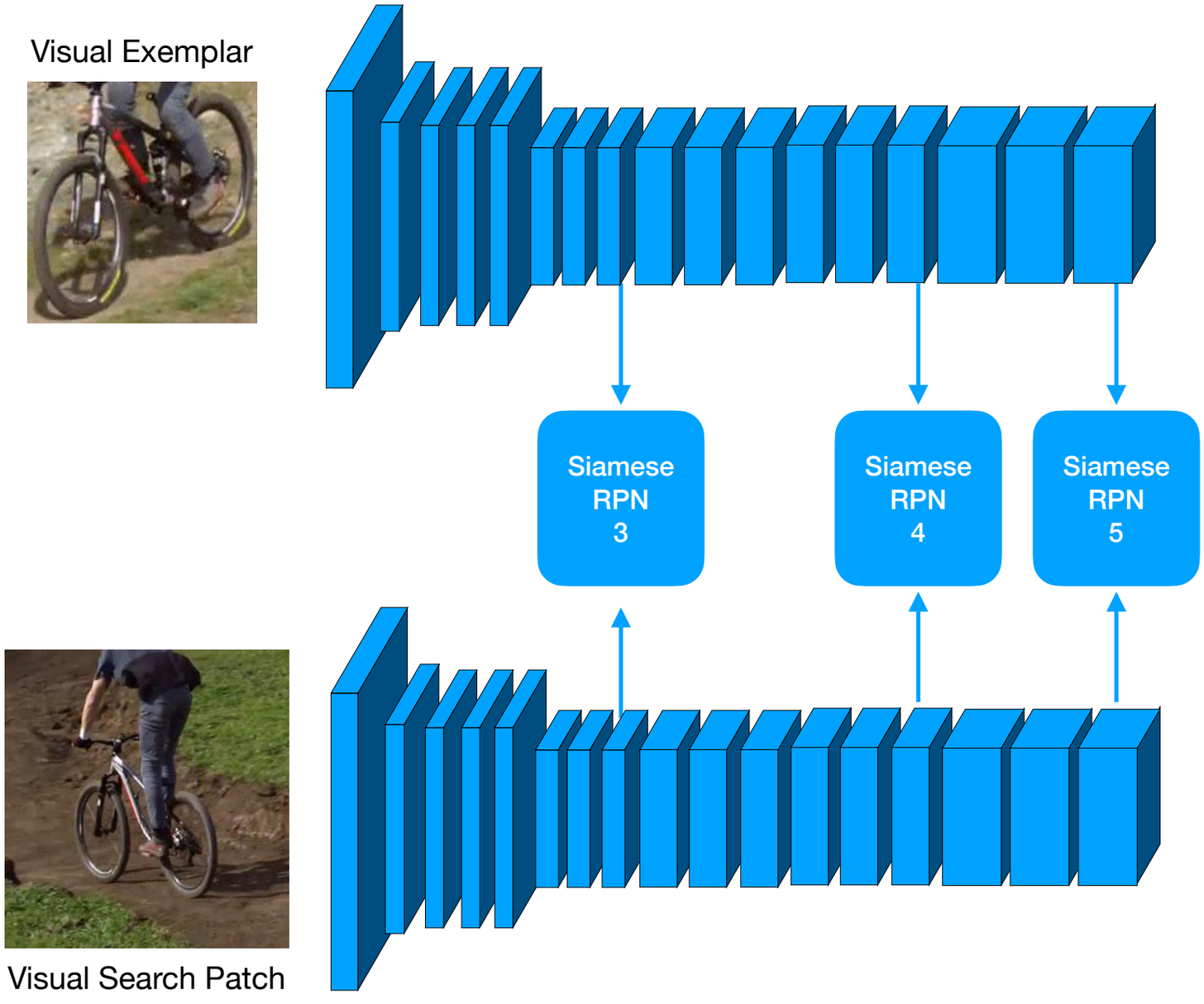

*Li, Bo, et al. "High performance visual tracking with siamese region proposal network." CVPR. 2018.*

# Siamese RPN



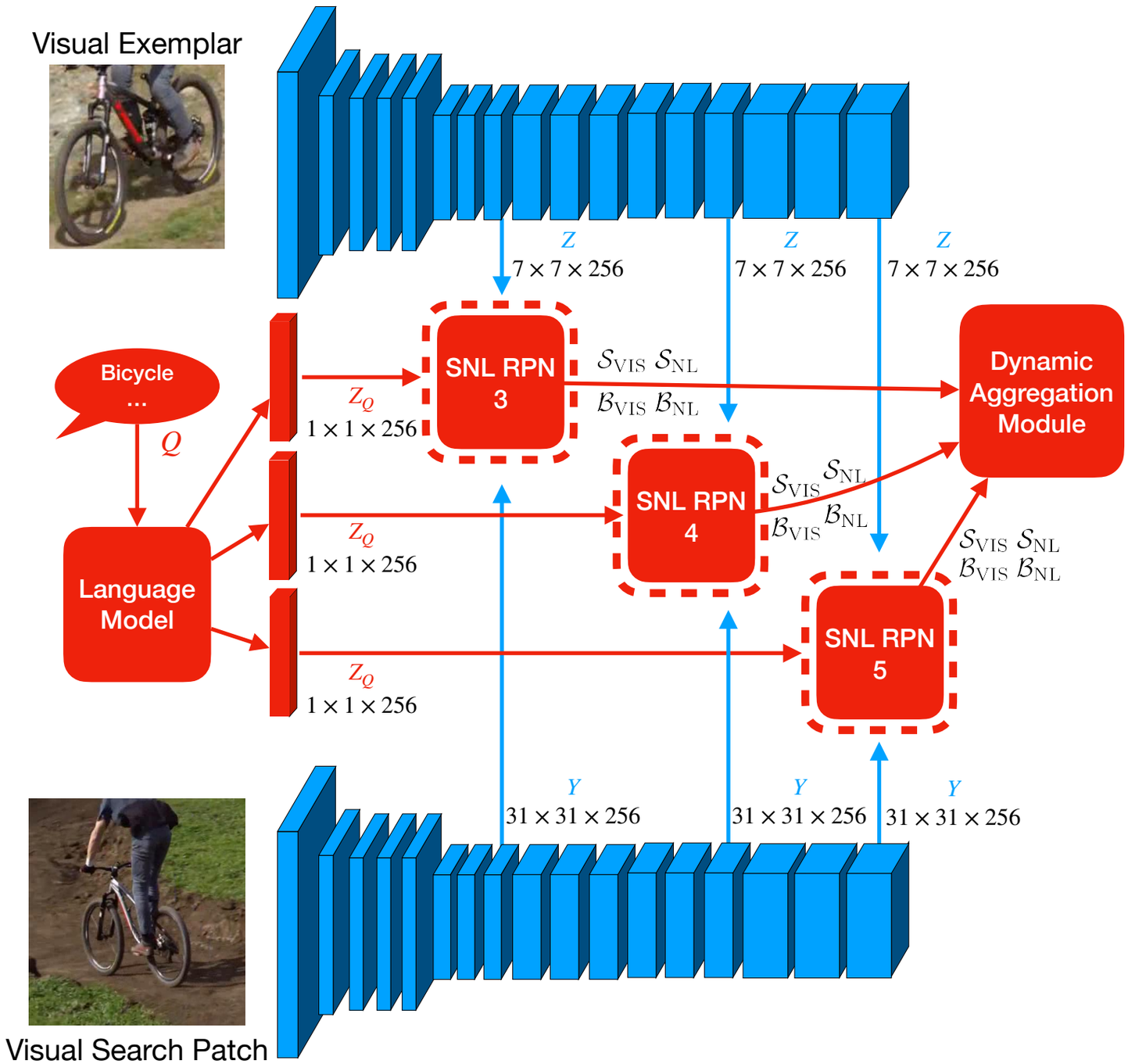

*Li, Bo, et al. "High performance visual tracking with siamese region proposal network." CVPR. 2018.*

# Siamese RPN++ [CVPR 19]



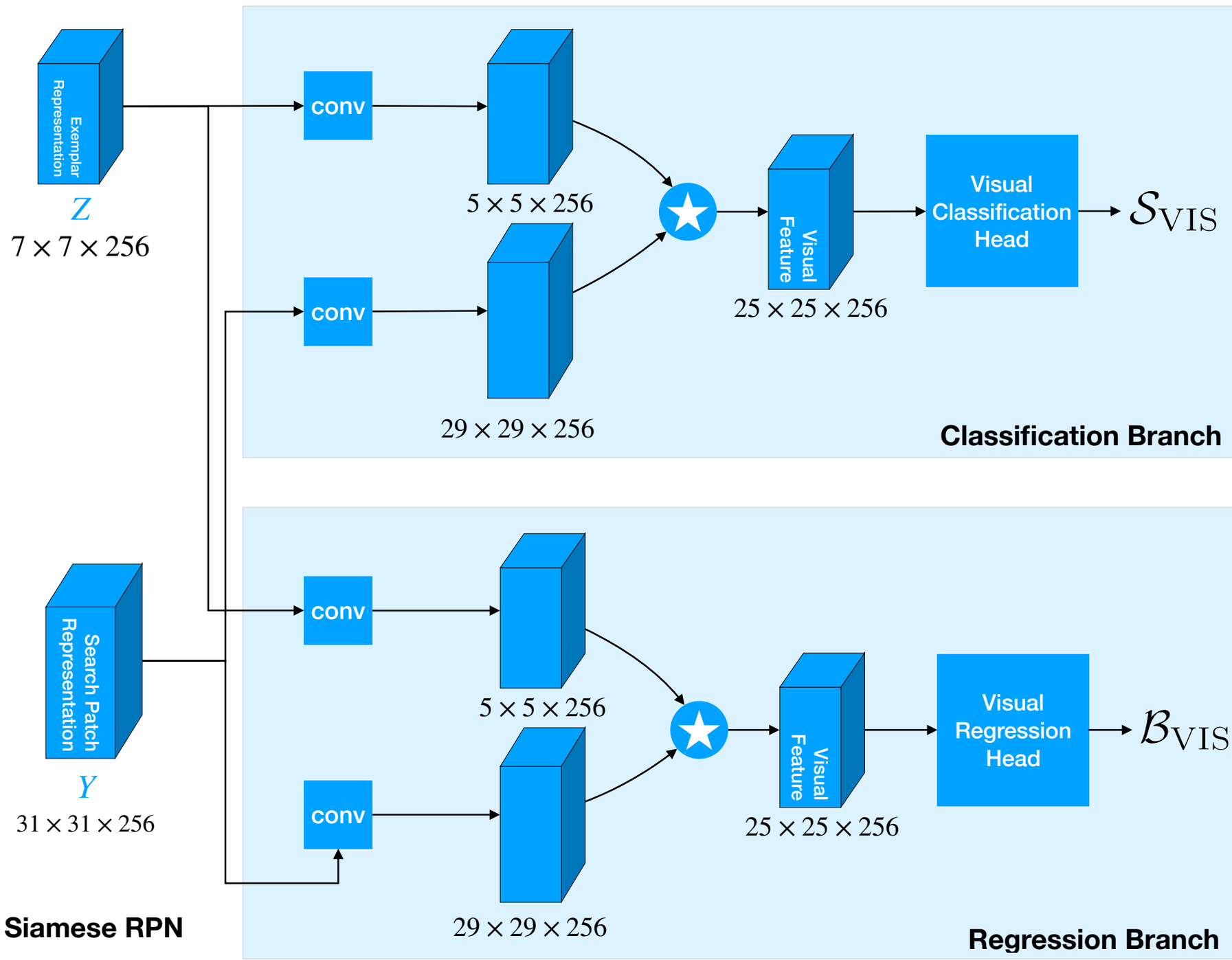

*Li, Bo, et al. "Siamrpn++: Evolution of siamese visual tracking with very deep networks." CVPR. 2019.*

# Siamese Natural Language Tracker

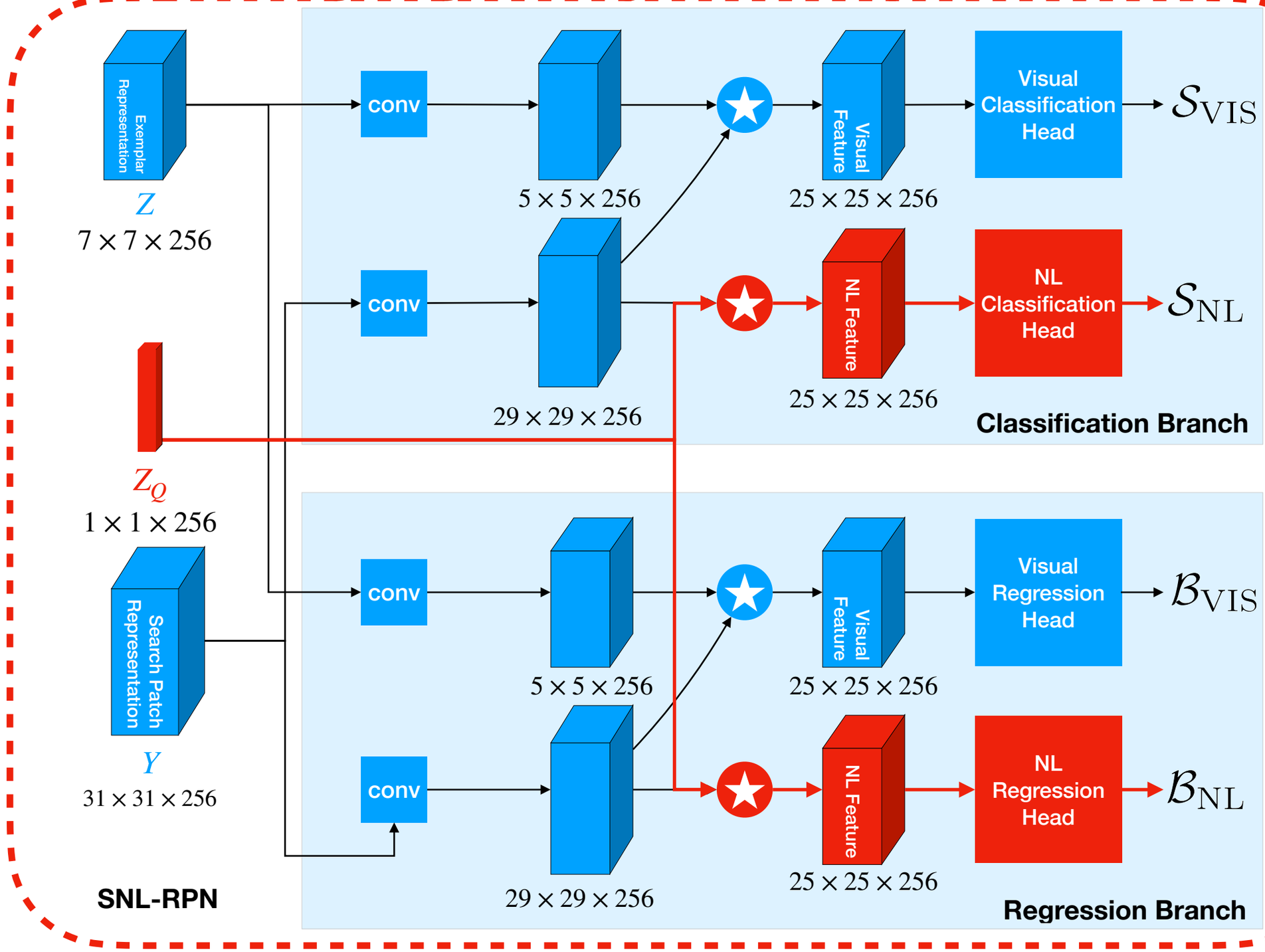



*Feng, Qi, et al. "Siamese Natural Language Tracker: Tracking by Natural Language Descriptions with Siamese Trackers." CVPR. 2021.*

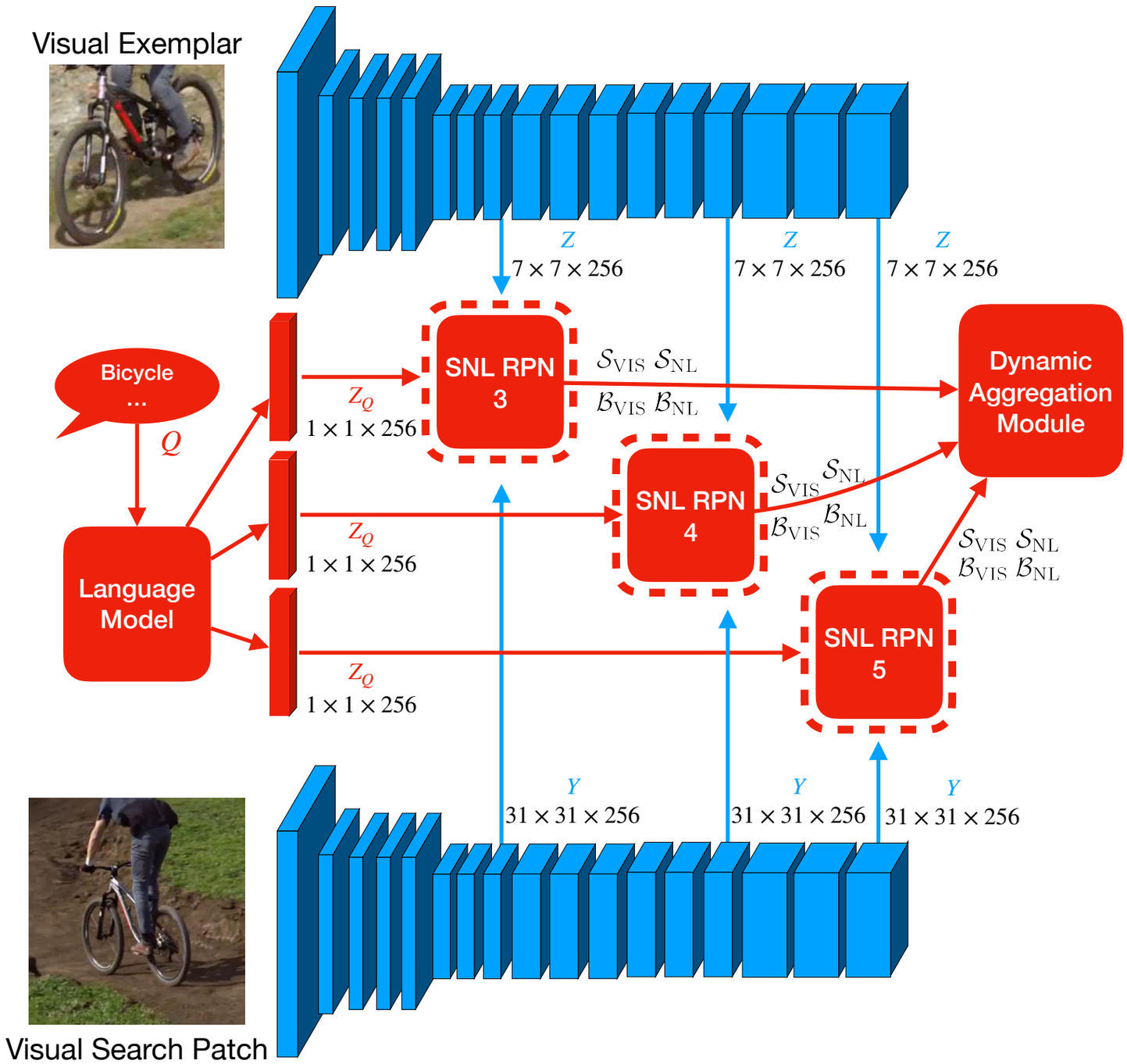
# Siamese RPN


Exemplar Representation
$Z$
$7 \times 7 \times 256$
conv
$5 \times 5 \times 256$
conv
$29 \times 29 \times 256$
Visual Feature
$25 \times 25 \times 256$
Visual Classification Head
$\mathcal{S}_{\mathrm{VIS}}$
Classification Branch
Search Patch Representation
$Y$
$31 \times 31 \times 256$
conv
$5 \times 5 \times 256$
conv
$29 \times 29 \times 256$
Visual Feature
$25 \times 25 \times 256$
Visual Regression Head
$\mathcal{B}_{\mathrm{VIS}}$
Regression Branch
Siamese RPN

# Natural Language Region Proposal Network



# Ours: Siamese Natural Language Tracker

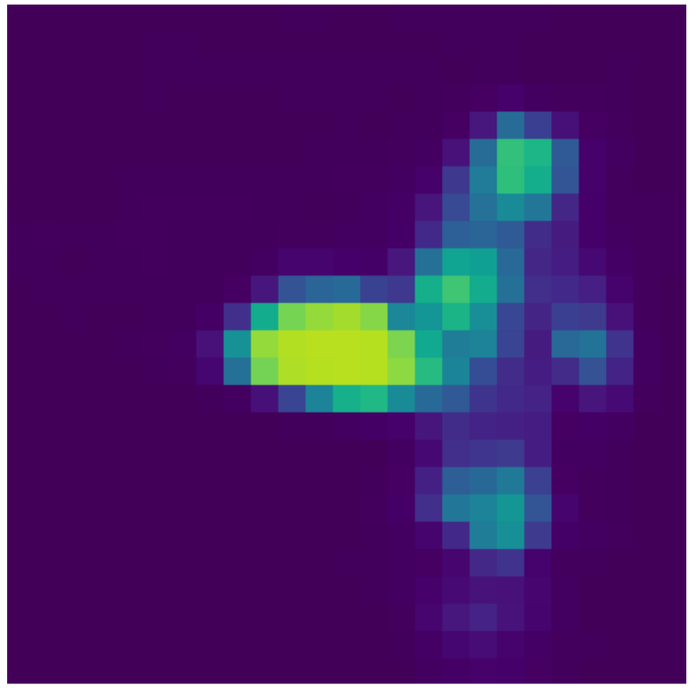



*Feng, Qi, et al. "Siamese Natural Language Tracker: Tracking by Natural Language Descriptions with Siamese Trackers" CVPR. 2021.*

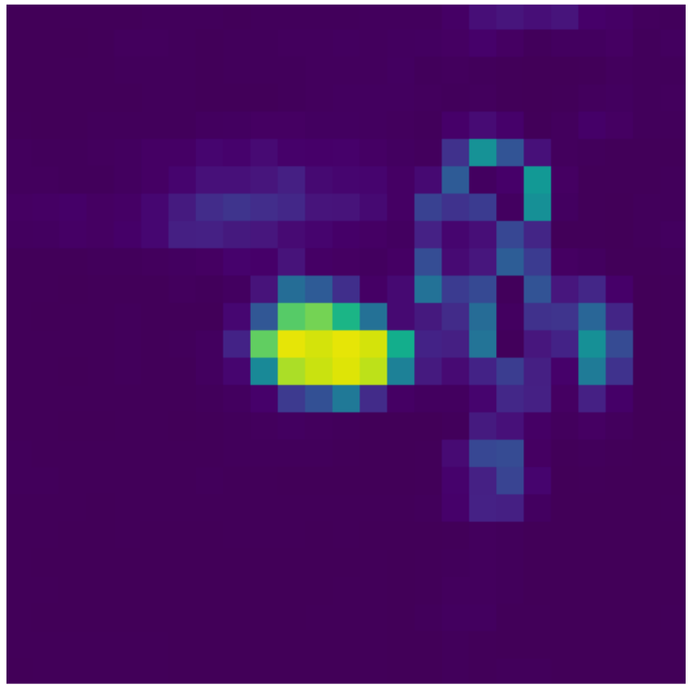
# Dynamic Aggregation

- Aggregate visual and language predictions by the entropies of the predicted heat map.
- Higher entropy → Lower weight.

Visual Search Patch

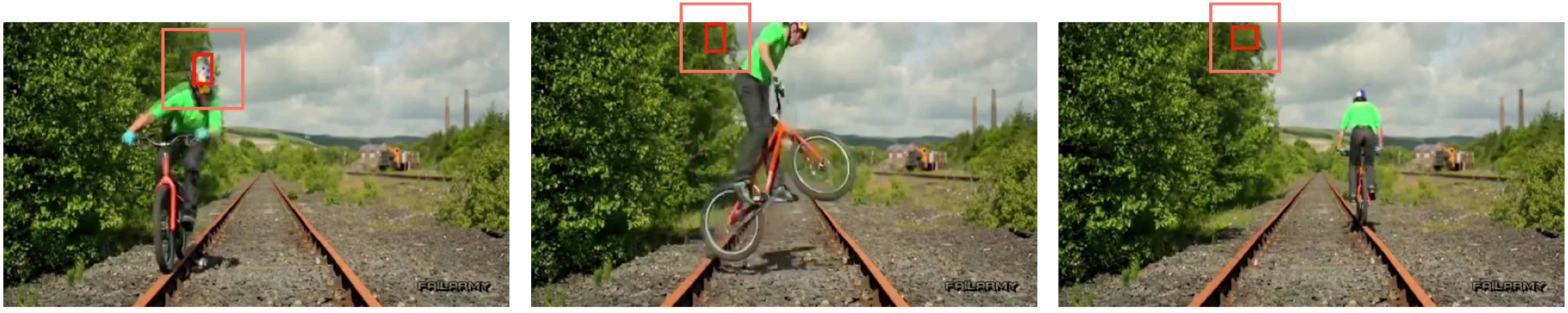


$\mathcal{S}_{\mathrm{VIS}}$

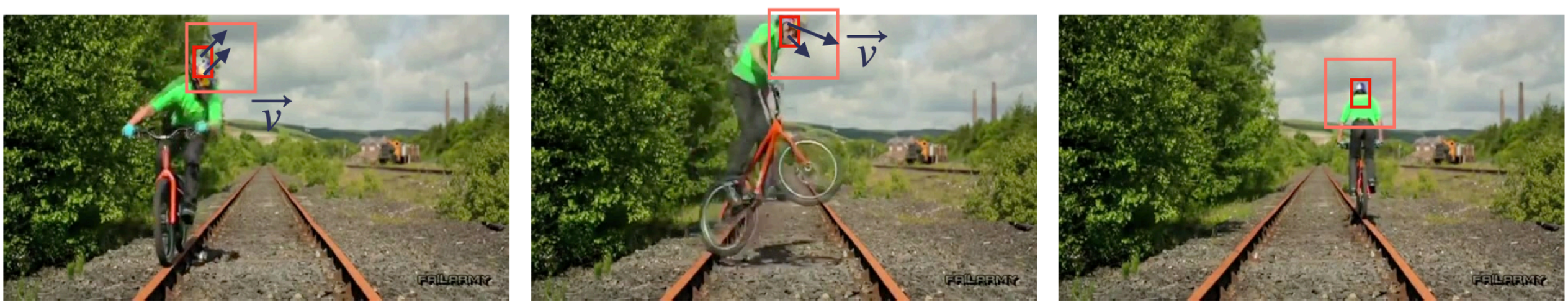


$\mathcal{S}_{\mathrm{NL}}$

# Optical Flow

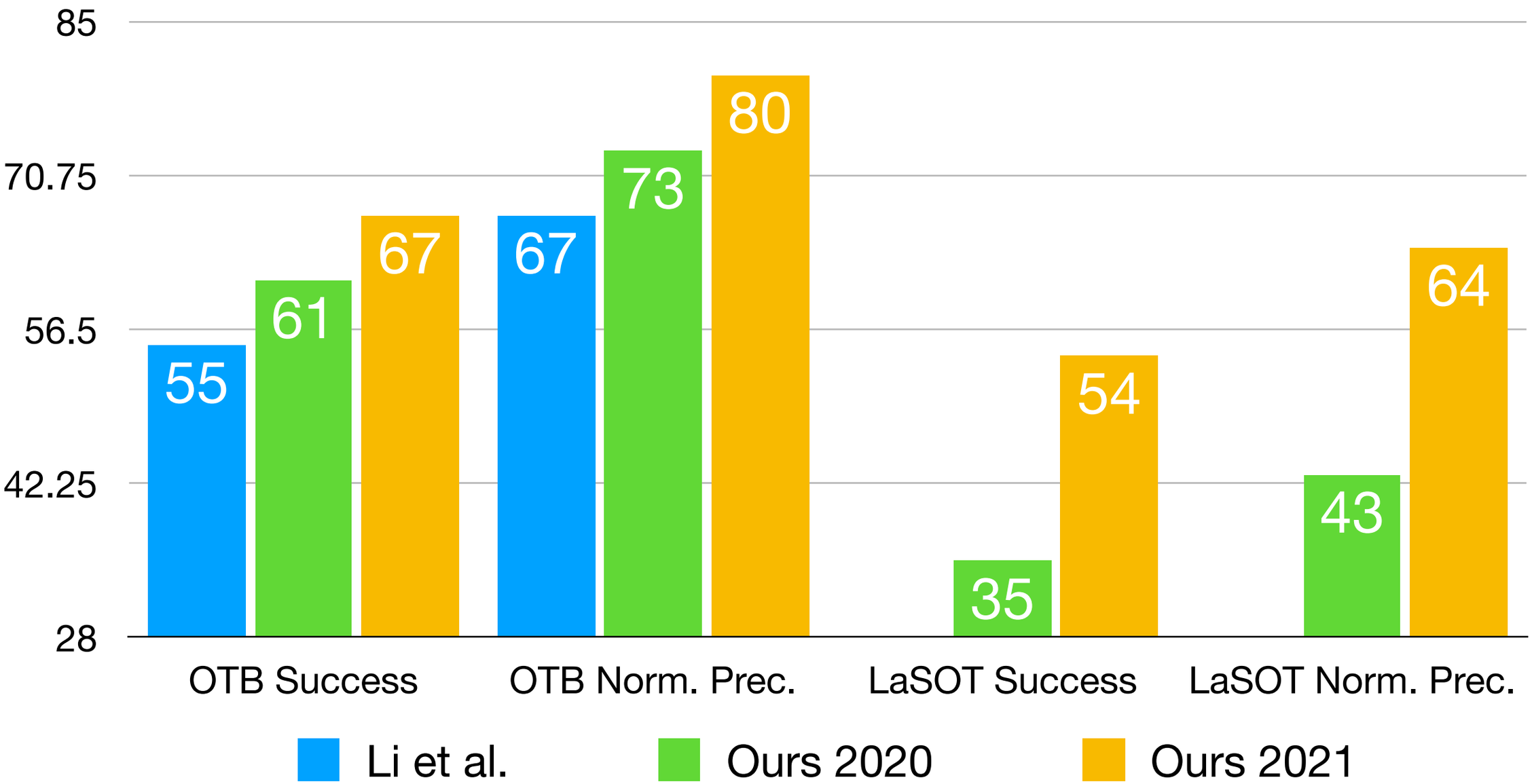


Without Optical Flow Guidance, targets that are moving rapidly will move outside of the search patch $X_t$.



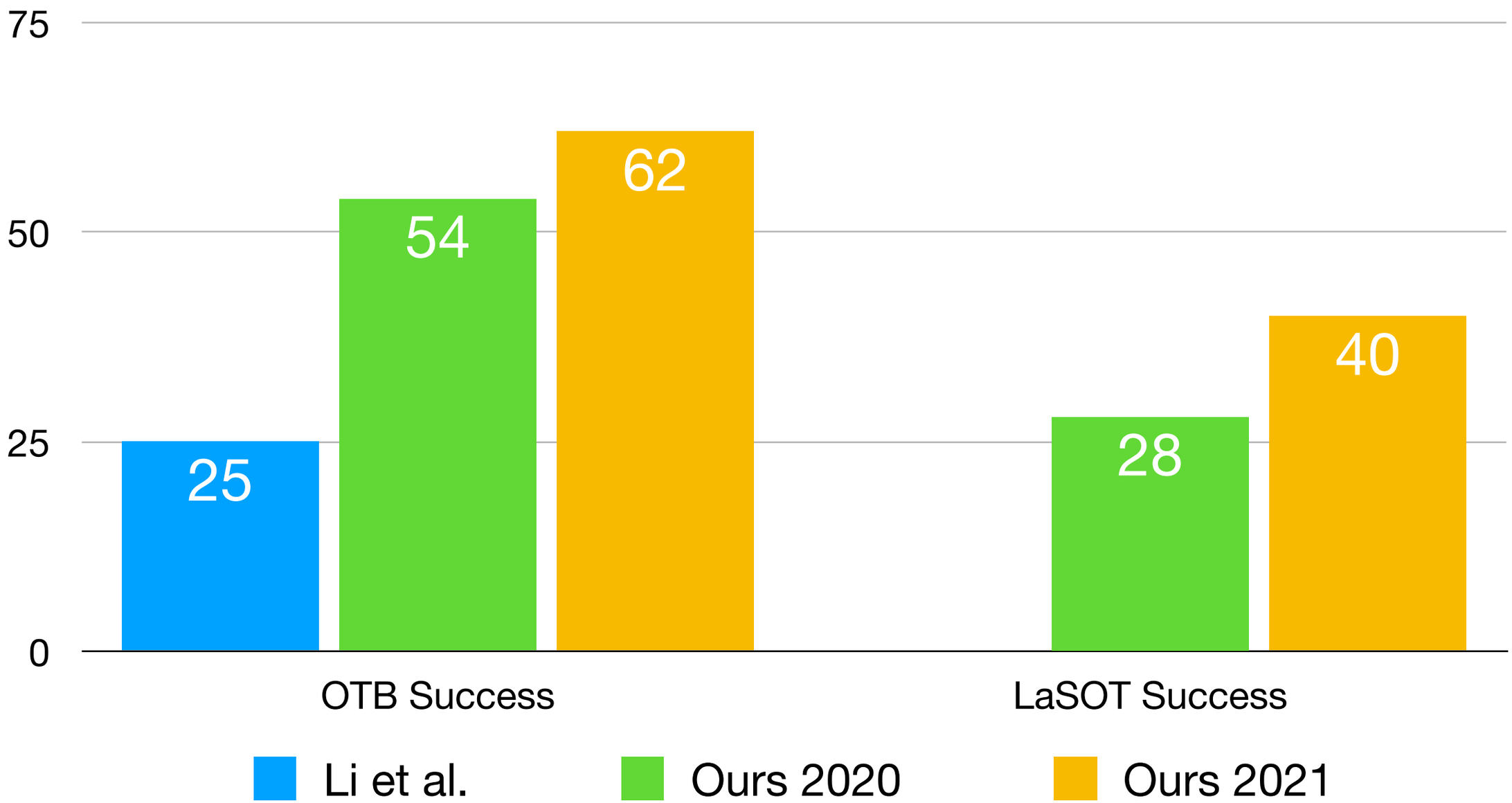

Using the average optical flow $\vec{v}$ within $\hat{B}_{t-1}$, the target is more likely to be covered by the search patch $X_t$.

# Results: Compared to prior work by Li *et al.*



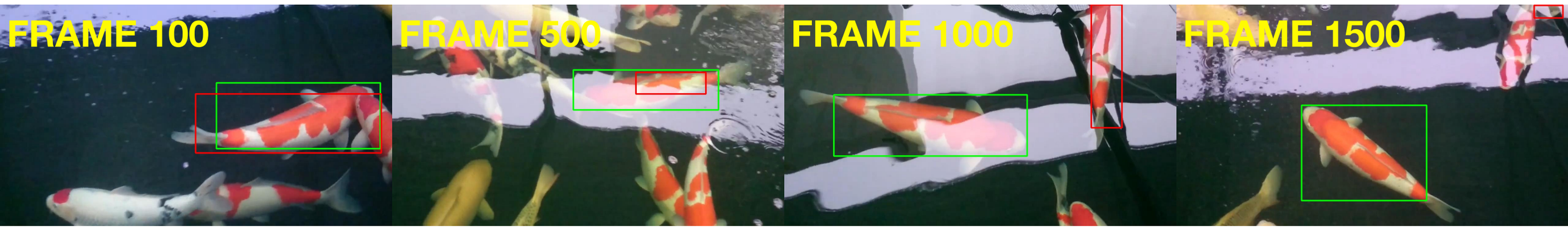

Performance of Li *et al.* tracker on LaSOT is not presented due to the lack of training codes.
Results obtained by using both bounding box and NL description as inputs.

# Results: Using Only NL Description for Tracking



# The Ambiguity of the NL Description $Q$


**The given NL uniquely describes the airplane and helps our tracker stay on the target, while SiamFC and SiamRPN++ suffer from model drifts.**

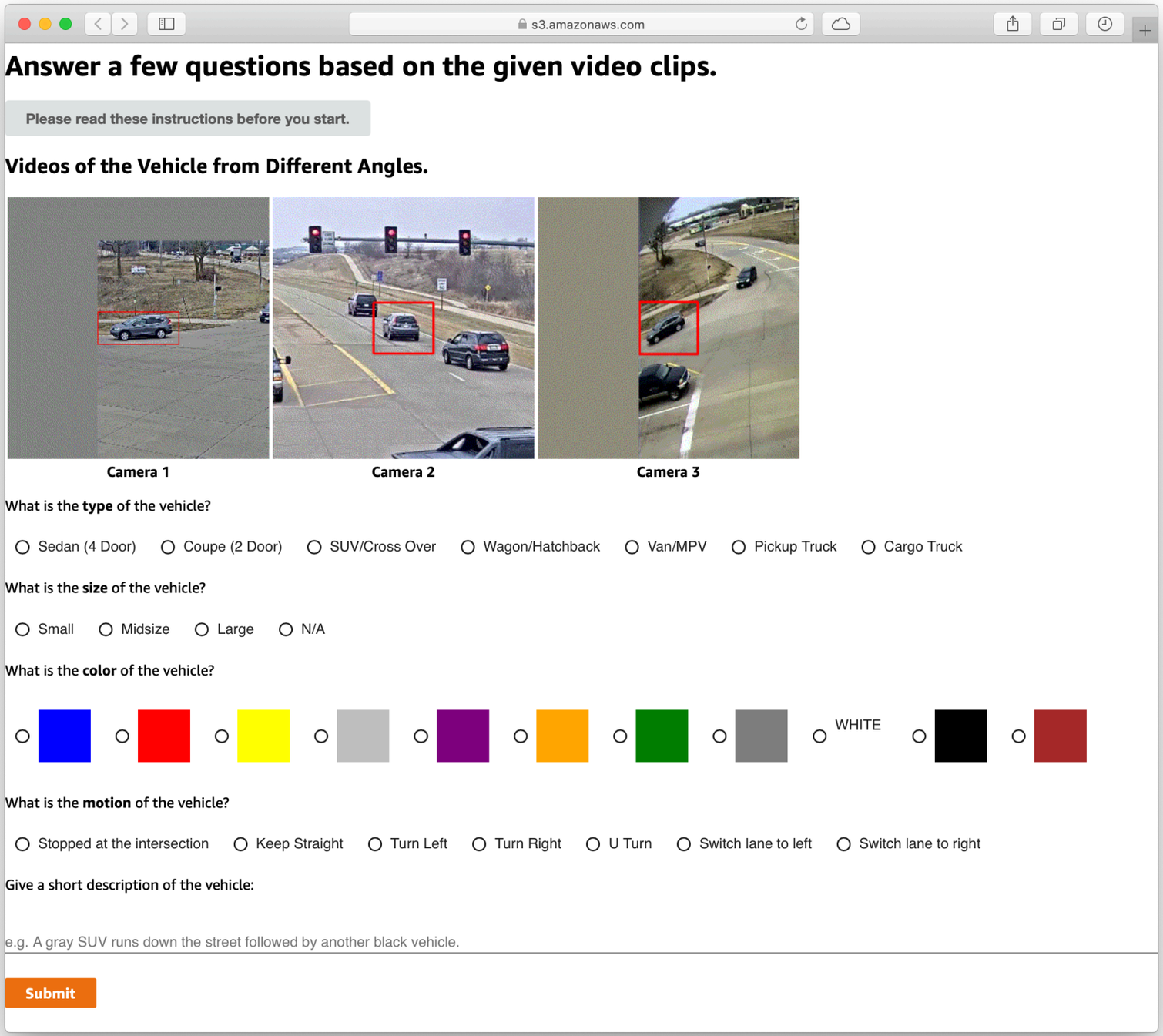



**The given NL (goldfish swimming among other fishes in the water) does not uniquely describe the target. As multiple goldfishes are present in the scene, the NL description does not help our tracker to avoid model drifting.**

# This Lecture

- Tracking By NL - single object - siamese approach
- CityFlow-NL Dataset
- Tracked-vehicle retrieval by NL

# Collecting NL Descriptions for Tracking



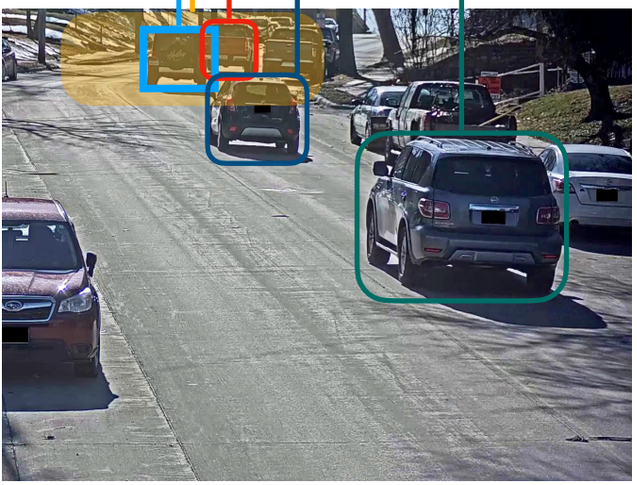

*Feng, Qi, et al. "CityFlow-NL: Tracking and Retrieval of Vehicles at City Scale by Natural Language Descriptions." arXiv:2101.04741.*

# CityFlow-NL

*A blue Jeep goes straight down a winding road behind a red truck and ahead of a black station wagon and black SUV.*

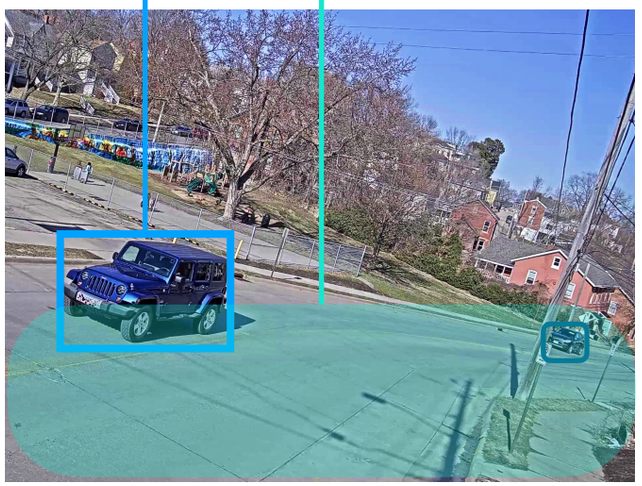
*A jeep going straight and then turning right on (the) road.*

*A blue truck running down the street straight behind a red pickup.*

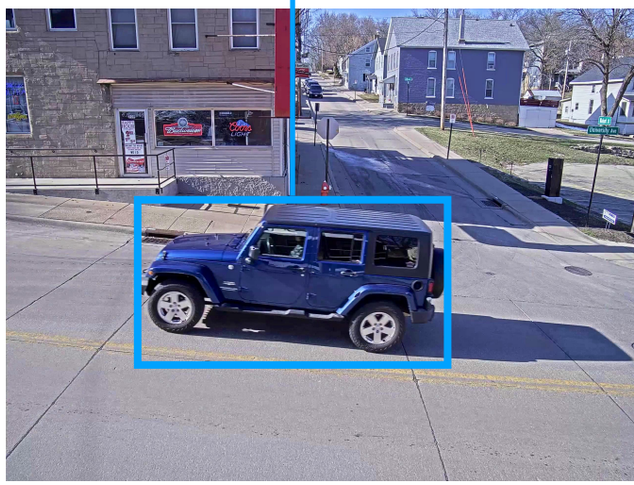


CAM 021 # 3877

CAM 020 # 3825



CAM 018 # 3783



CAM 019 # 3837

*Feng, Qi, et al. "CityFlow-NL: Tracking and Retrieval of Vehicles at City Scale by Natural Language Descriptions." arXiv:2101.04741.*

# CityFlow-NL: Another Example

**NL Descriptions for Track 28**

A white SUV drives down the road and passes parked cars behind a sliver sedan.

A white Crossover approaches a stop sign and is second in line behind a gray van.

A white SUV goes straight down a street and comes to a stop behind another silver vehicle.

A silver sedan is driving straight down a road followed by a white SUV.

A small gray sedan turns left followed by another white vehicle.

Gray sedan goes straight followed by a white vehicle.

*Conflicting descriptions due to different perspective.*

**NL Descriptions for Track 136**

*Feng, Qi, et al. "CityFlow-NL: Tracking and Retrieval of Vehicles at City Scale by Natural Language Descriptions." arXiv:2101.04741.*

# This Lecture

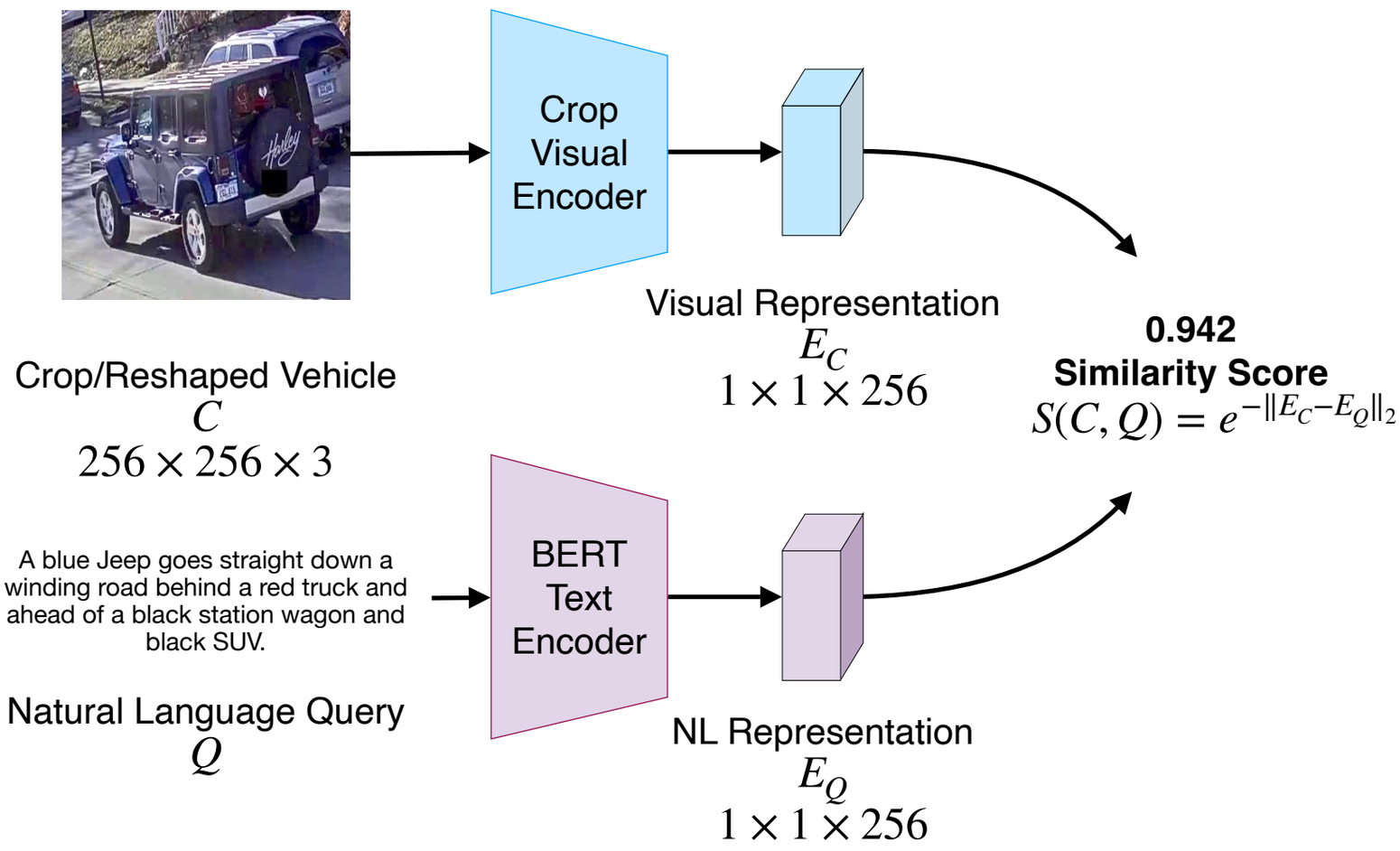
- Tracking By NL - single object - siamese approach
- CityFlow-NL Dataset
- Tracked-vehicle retrieval by NL

# A simpler task: tracked-object retrieval by NL

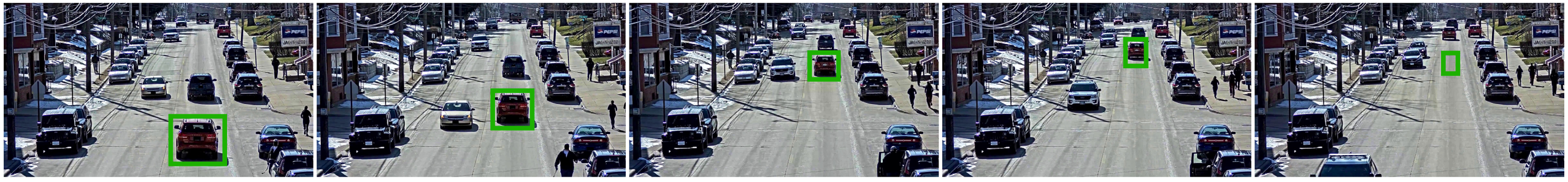
- https://www.aicitychallenge.org
- Natural Language-Based Tracked Vehicle Retrieval.
- Given tracked vehicles and an NL query $Q$.
- Goal: Rank all candidate vehicle tracks by the query.
- Evaluated with 530 test queries and tracks using MRR.
  - $$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}.$$
  - $|Q| = 530$ is the size of the test set.

*Naphade, Milind, et al. "The 5th AI City Challenge" CVPRW. 2021.*

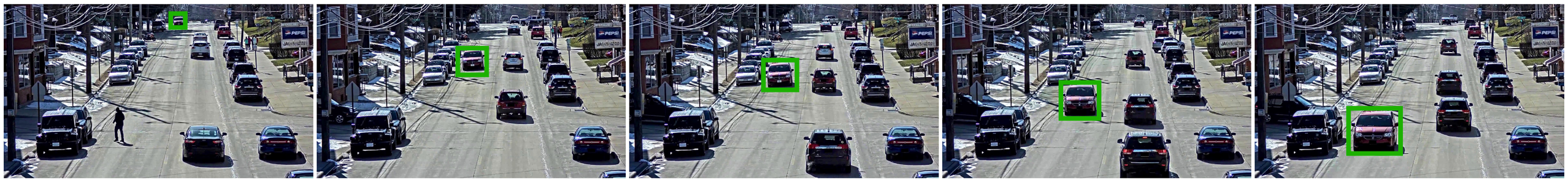
# Baseline Model



# Failure Case



The NL descriptions given are: “A red wagon goes straight”, “Red SUV going straight in the right lane.” and “A red SUV head straight down the road.” The baseline retrieval model gives this track the average similarity score at 0.81 and ranks this track the fifth. Notice that the second NL description specified that the target is moving in the right lane.



The baseline retrieval model ranks this track the first with the highest average similarity score at 0.84. This vehicle shares the same size, type, color, and motion pattern with the target vehicle for retrieval. The baseline retrieval model fails to capture the spatial referring expression (”in the right lane”) from the second NL description.

# Note

- We only considered crop visual features in the baseline model.
- Vehicle motion / Background information / Relations to other vehicles are ignored in this approach.

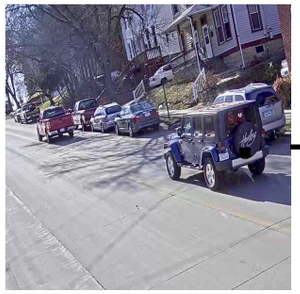
- We released the pre-processed annotations to the public.
- A quad-stream model is then proposed based on public response.

# Quad-stream Model

A blue Jeep goes straight down a winding road behind a red truck and ahead of a black station wagon and black SUV.

CLIP Text Encoder

RELATION HEAD | CROP HEAD | CLIP HEAD | MOTION HEAD | JOINT HEAD



Reshaped Frame
$F$

Relation Visual Encoder

CLS & XE Loss

Relation NL HEAD

**0.94 Relation Similarity**

InfoNCE

Crop/Reshaped Vehicle
$C$

Crop Visual Encoder

CLS & XE Loss

CROP NL HEAD

**0.92 Crop Similarity**

InfoNCE

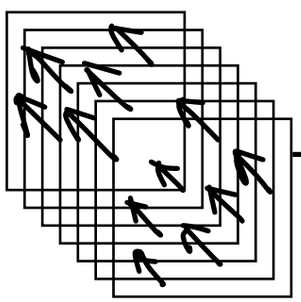
CLIP Visual Encoder

CLS & XE Loss

CLIP NL HEAD

**0.99 CLIP Similarity**

InfoNCE



Vehicle Optical Flows
$O$

Motion Encoder (GRU)

CLS & XE Loss

MOTION NL HEAD

**0.85 Motion Similarity**

InfoNCE

WEIGHTED AVERAGE

**0.93 Overall Similarity**

InfoNCE

CLS & XE Loss

Concat & Joint

**0.95 Joint Similarity**

InfoNCE

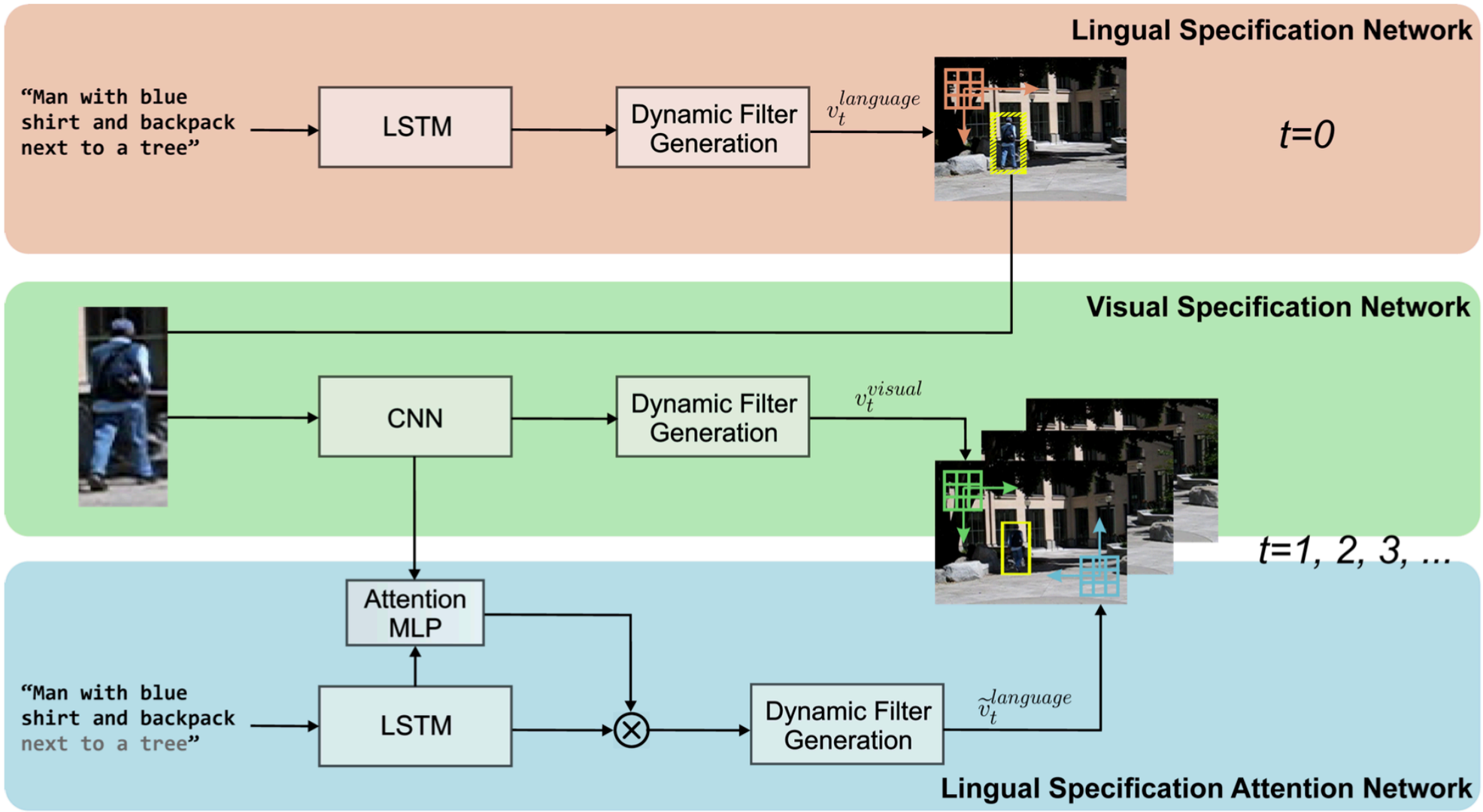
# Retrieval Performance

| Model | Visual Feature | Motion | Relation | NL Embedding | MRR |
|---|---|---|---|---|---|
| **Alibaba-UTS-ZJU** | SE-ResNet 50 | ✓ | ✓ | BERT | 0.1869 |
| **SDU-XidianU** | ResNet101-ibn-a | ✗ | ✗ | GloVe | 0.1613 |
| **SUNYKorea** | ResNet-50 | ✓ | ✓ | Part-of-speech | 0.1594 |
| **Sun Asterisk** | CLIP | ✗ | ✗ | CLIP | 0.1571 |
| **Quad-stream** | CLIP & SE-ResNet 50 | ✓ | ✓ | CLIP | 0.2294 |
| **Baseline** | ResNet-50 | ✗ | ✗ | BERT | 0.0269 |

*Naphade, Milind, et al. “The 5th AI City Challenge” CVPRW. 2021.*
*Feng, Qi, et al. "CityFlow-NL: Tracking and Retrieval of Vehicles at City Scale by Natural Language Descriptions." arXiv:2101.04741.*

# Additional Slides

# Prior work by Li *et al.*




*Li, Zhenyang, et al. "Tracking by natural language specification." CVPR. 2017.*