# Scene Graph Generation from Objects, Phrases and Region Captions

Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, Xiaogang Wang

발표자 : 한연지

# 논문 소개

- ICCV 2017 paper

(ECCV 2018에도 기본 구조로도 제출됨)

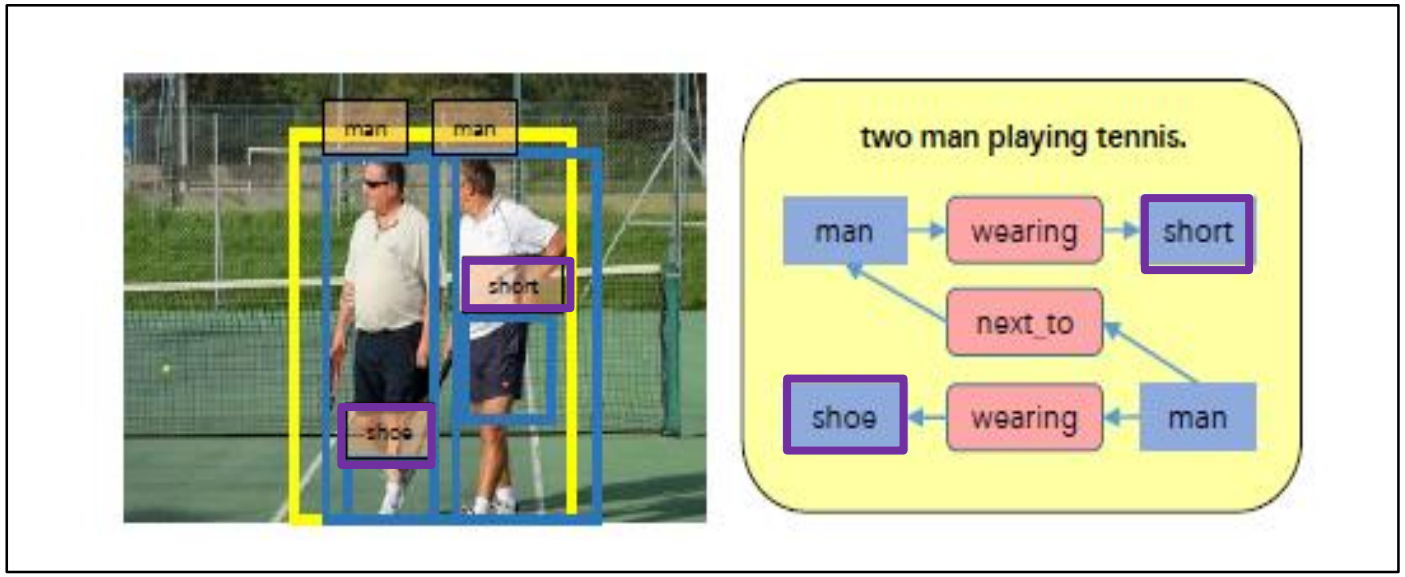
- LI Yikang's Interests

- Computer Vision
- Robotics

<한 줄 요약 : 이미지 설명할 수 있는(객체, 동사 등을 detection) 여러 모델을 하나의 모델로 사용하여 지식 그래프를 완성>

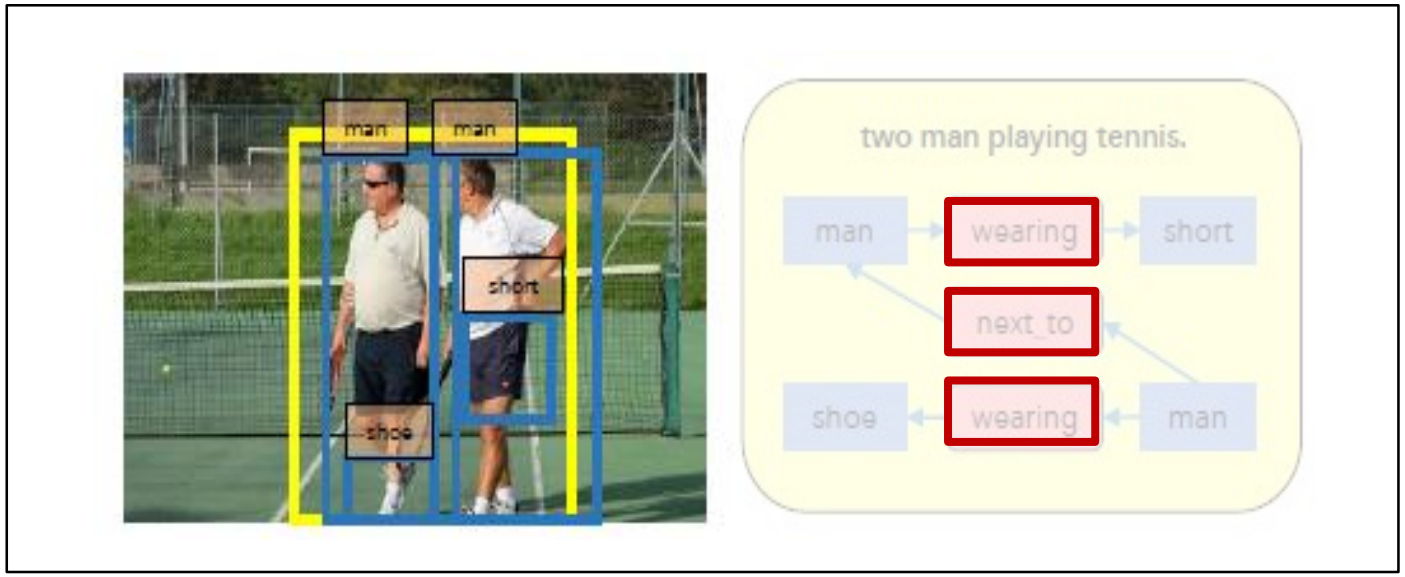
LI Yikang

# Conventional Approach




Object Detection

# Conventional Approach




**Predicate Detection**

# Conventional Approach

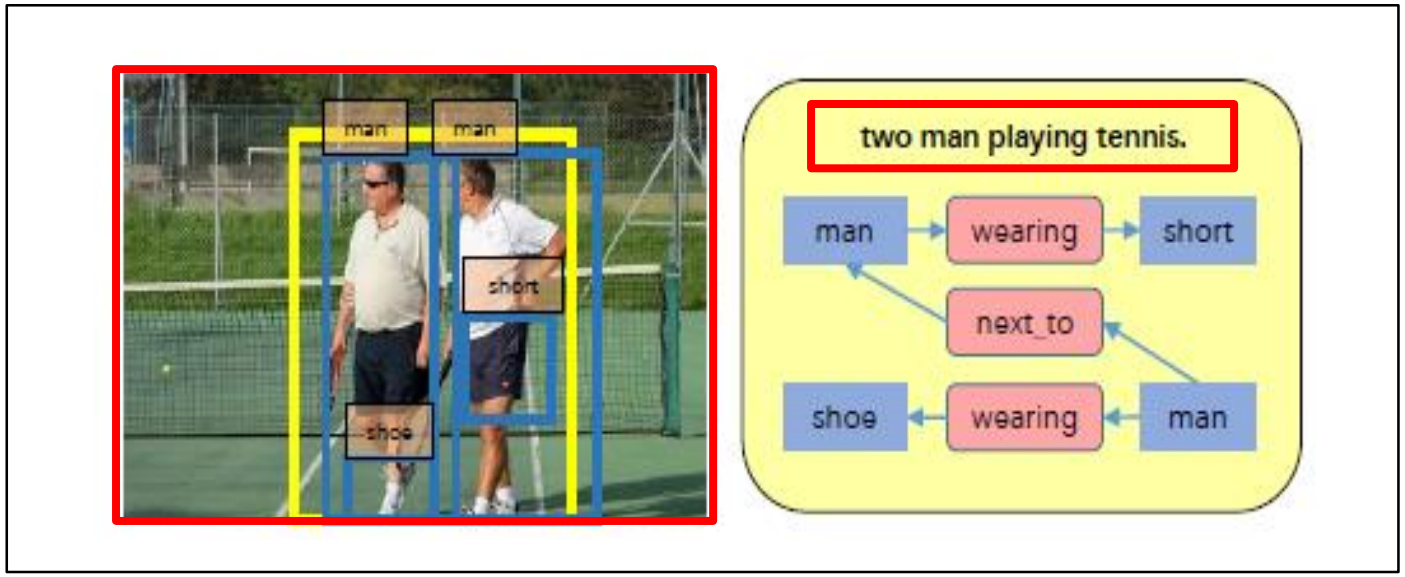
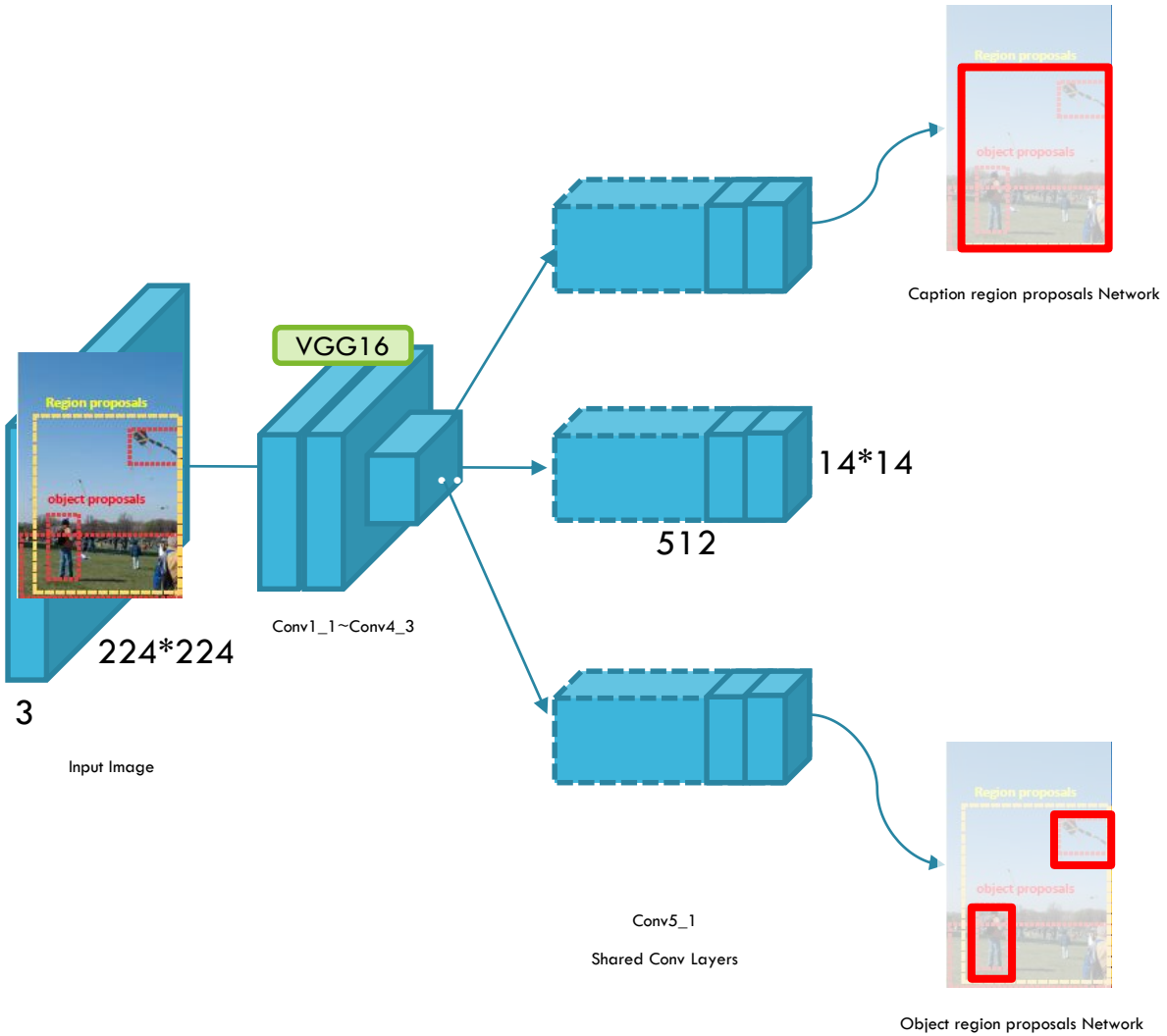
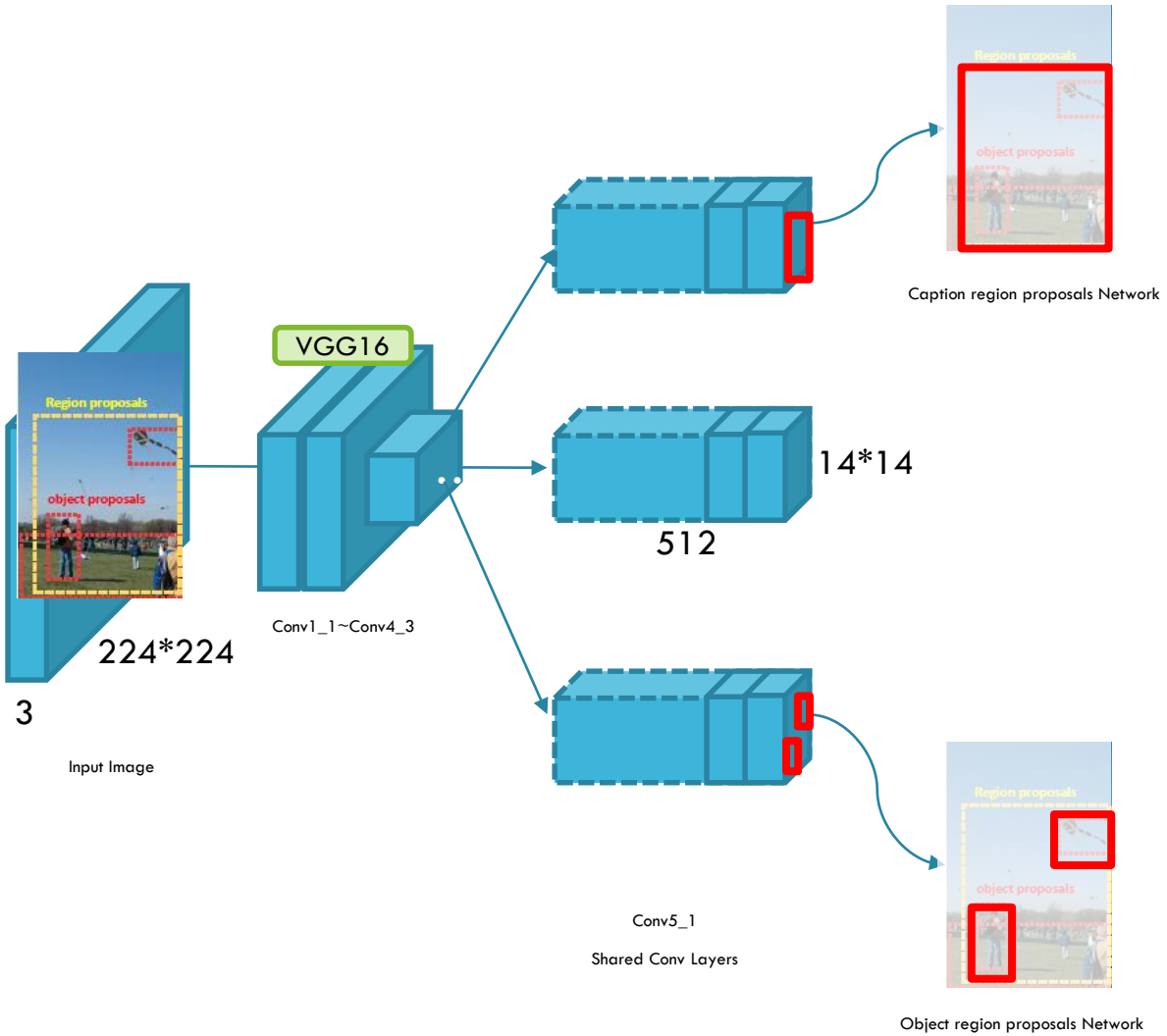



Image Caption

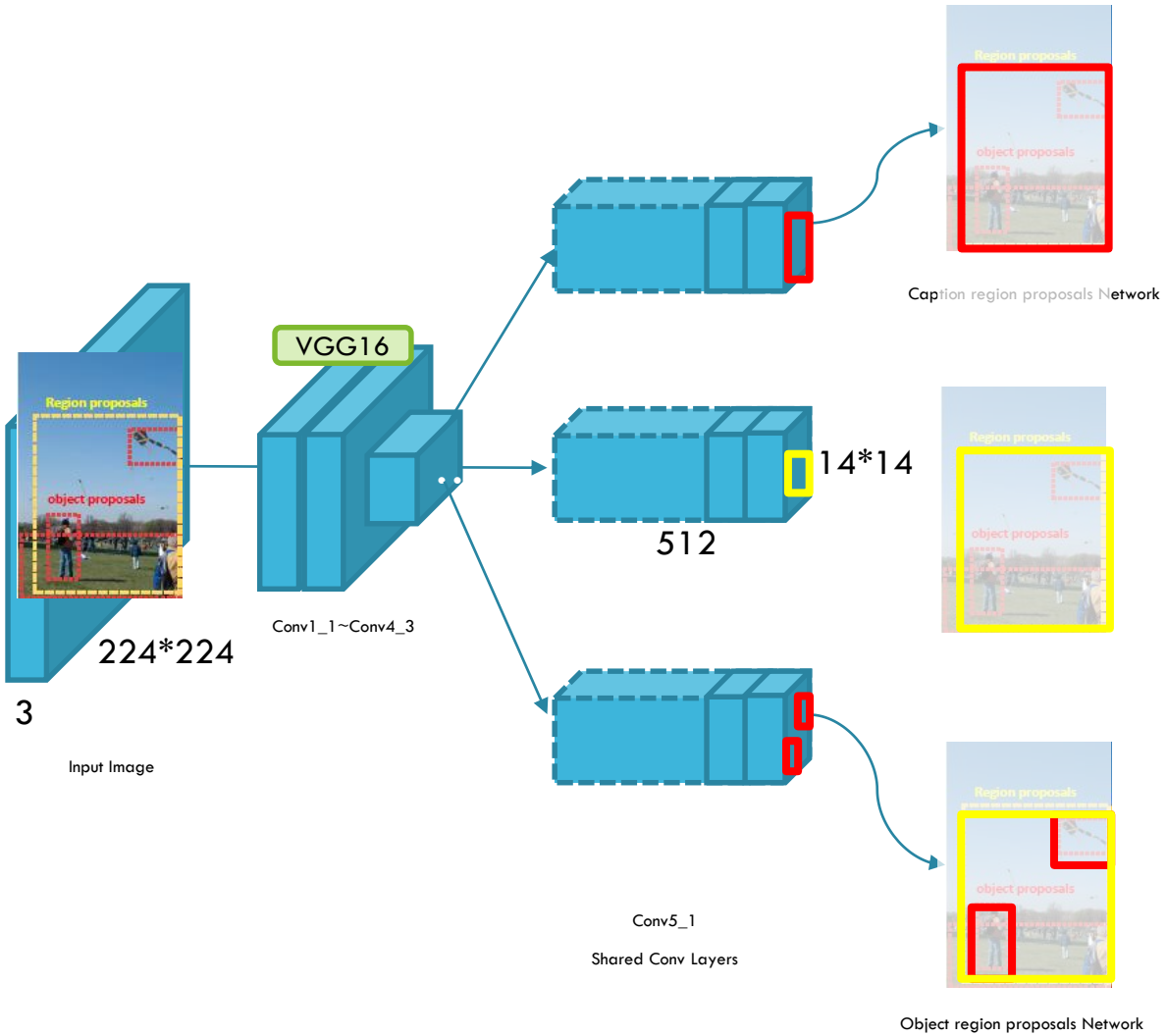
# Multi-level Scene Description Network



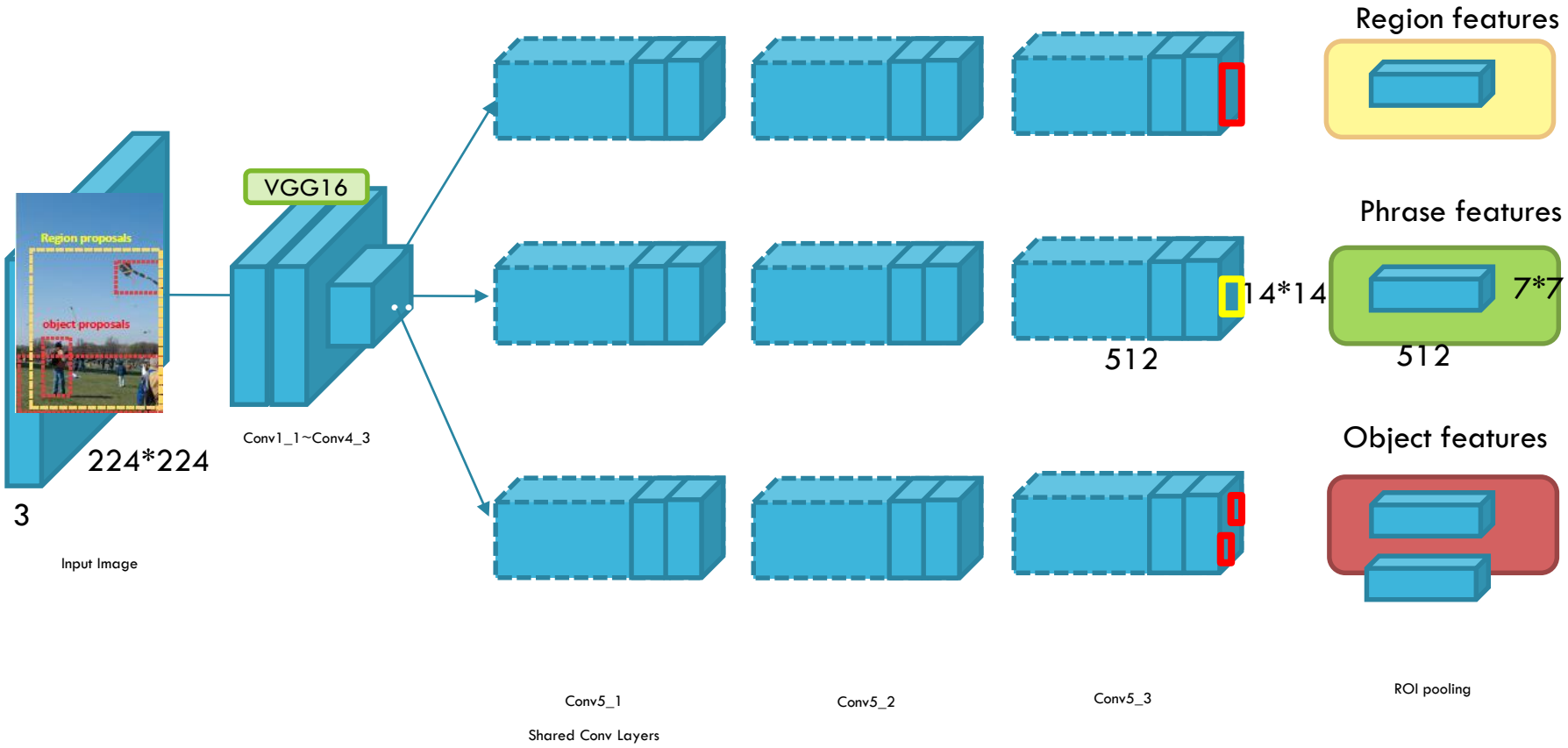
# Multi-level Scene Description Network



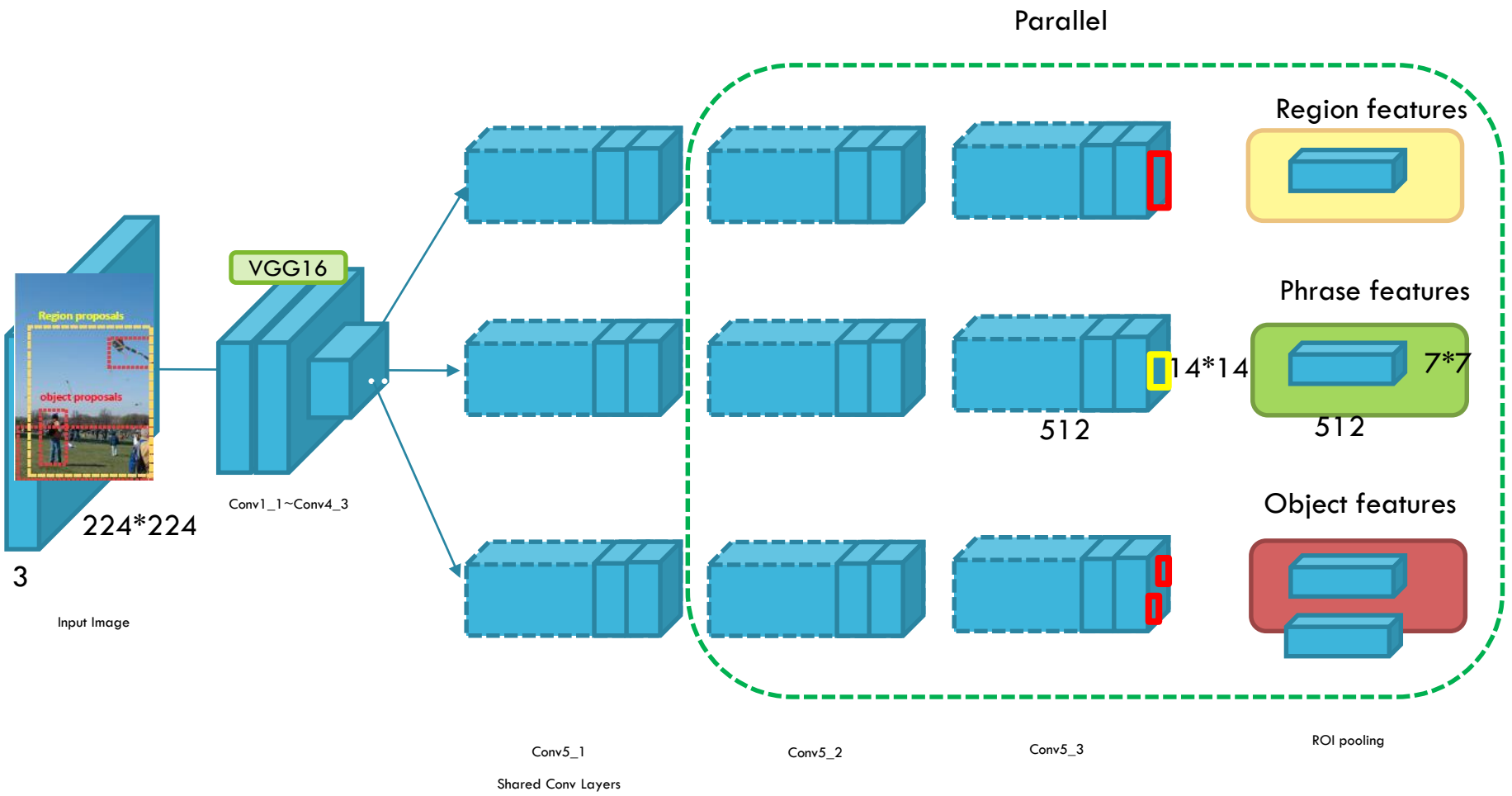
# Multi-level Scene Description Network



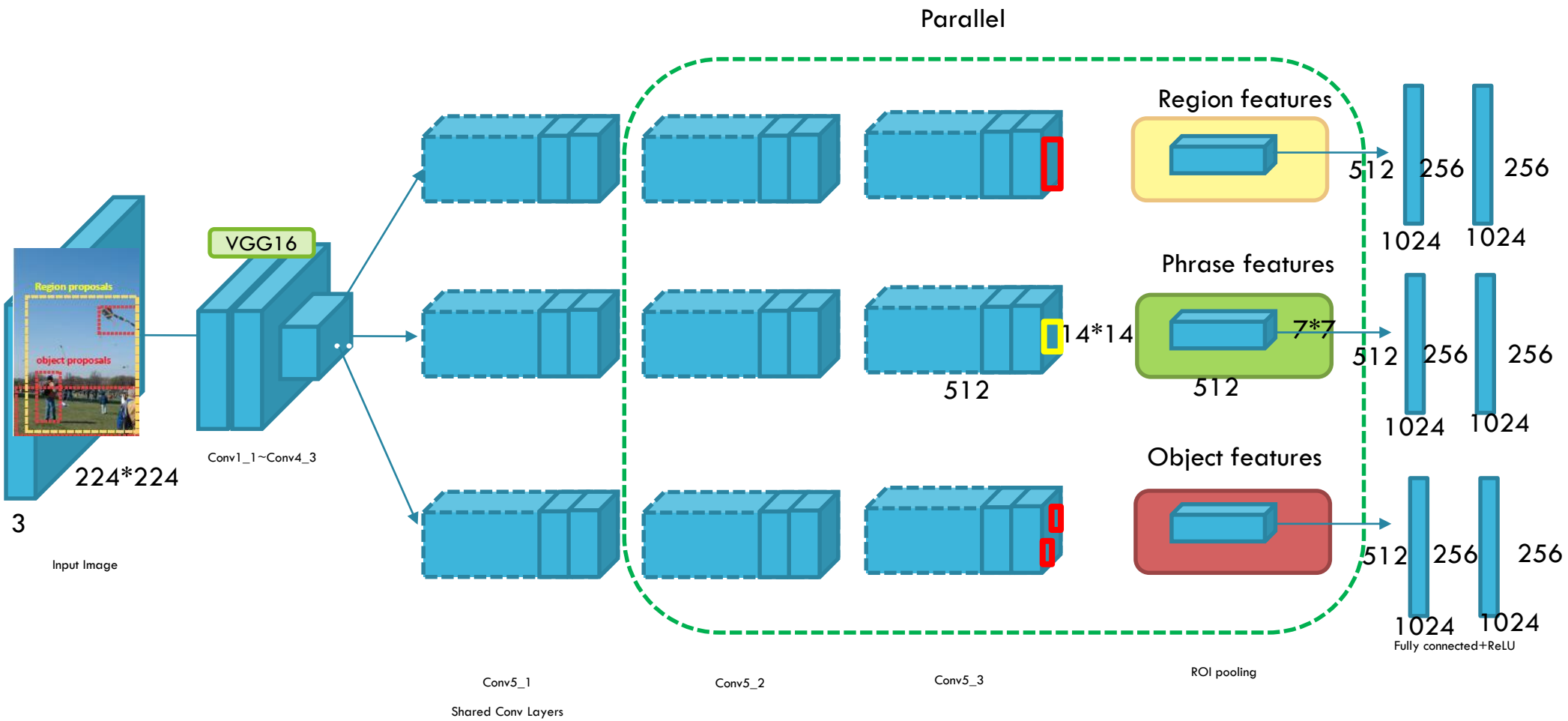
# Multi-level Scene Description Network



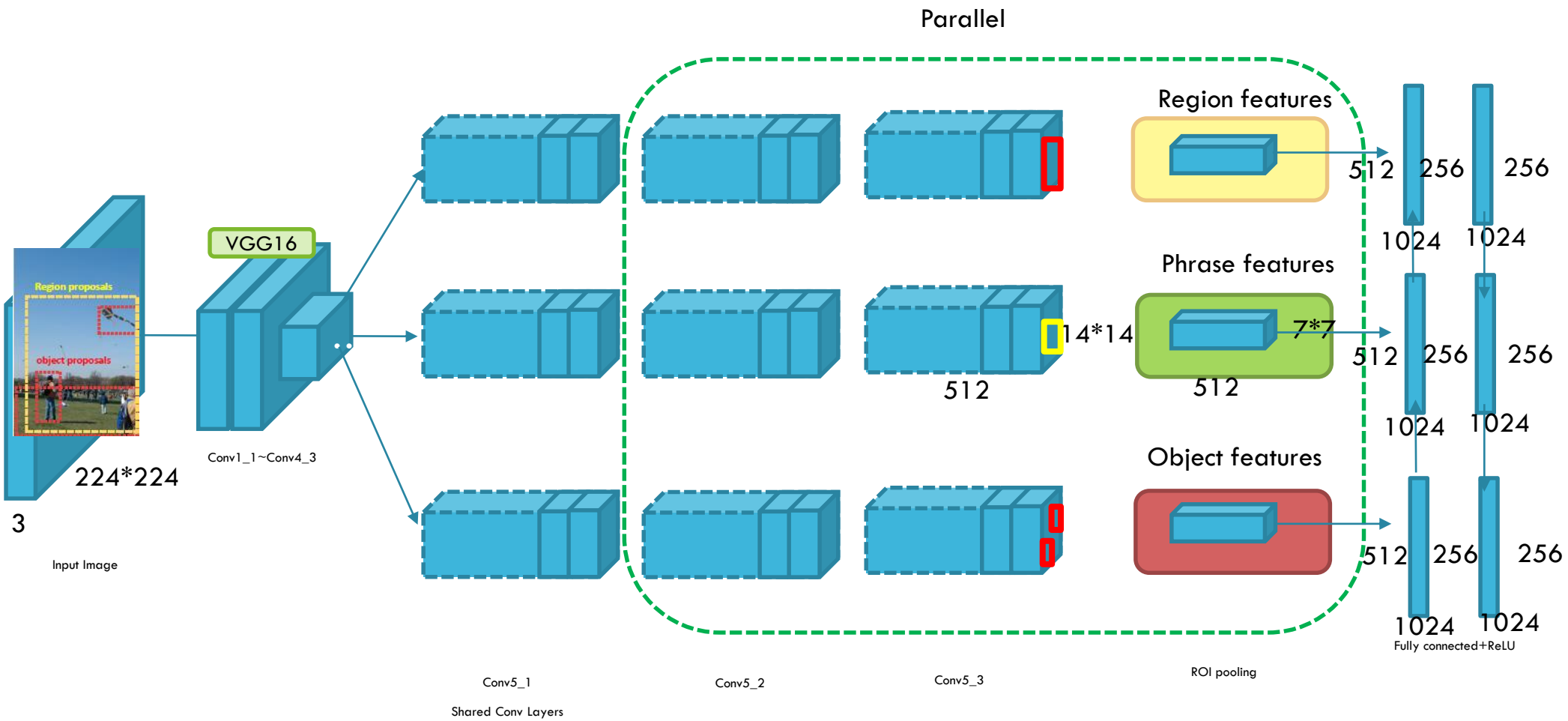
# Multi-level Scene Description Network



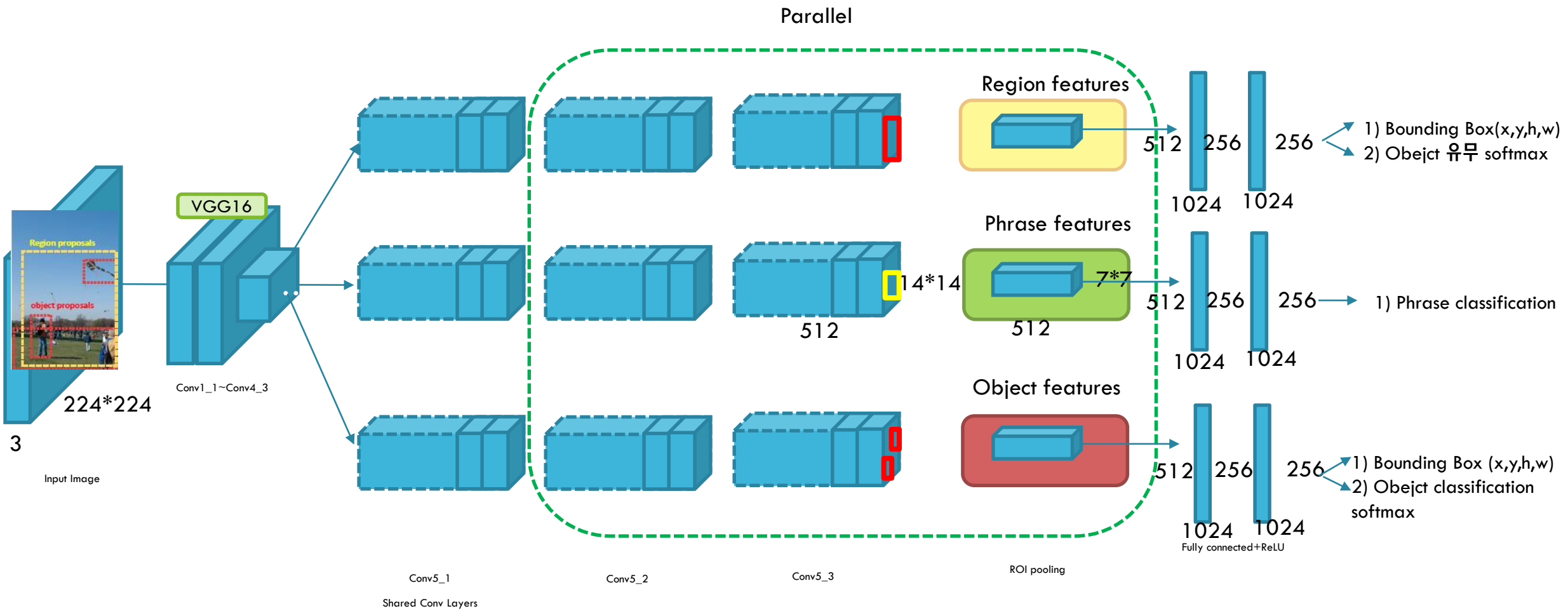
# Multi-level Scene Description Network



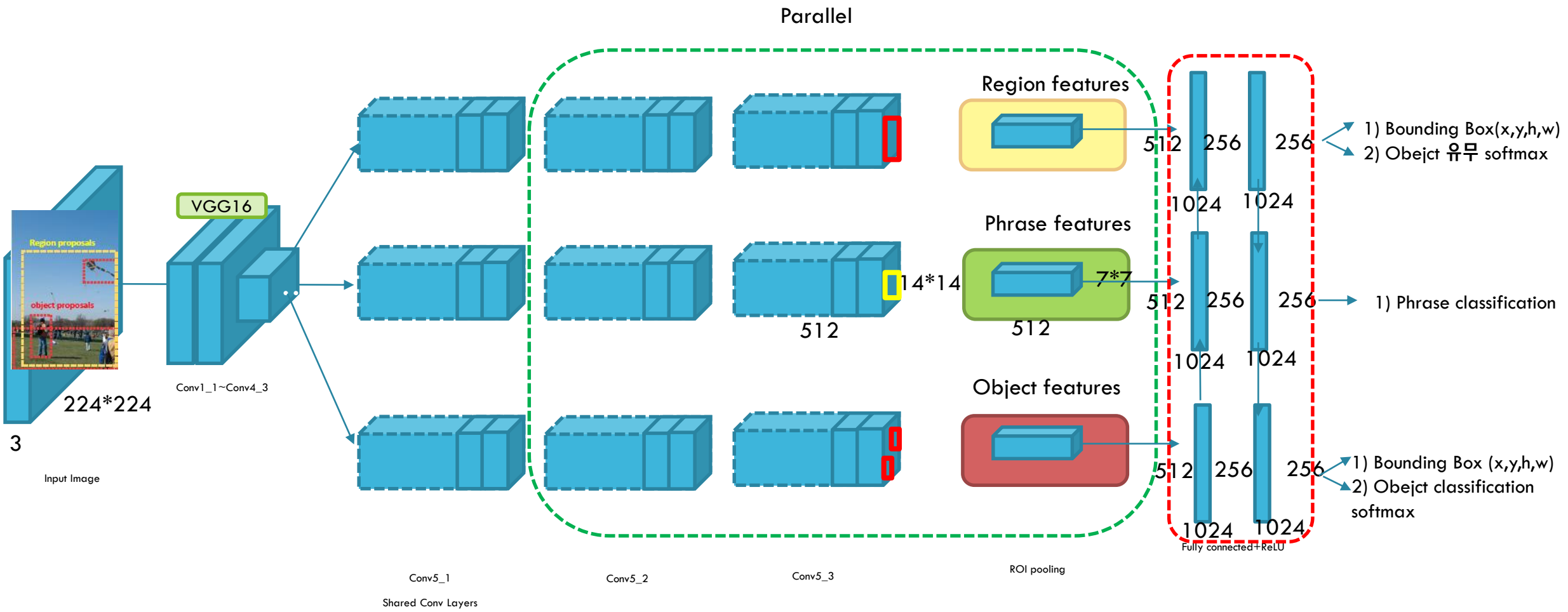
# Multi-level Scene Description Network



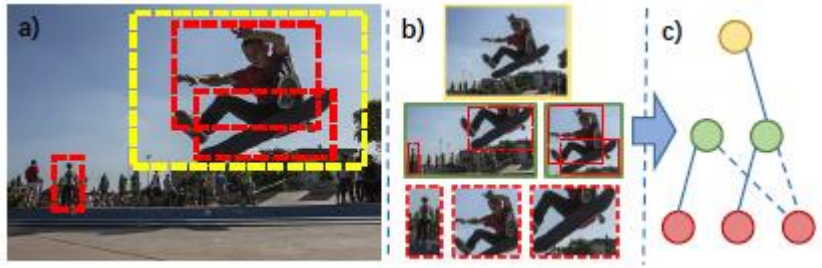
# Multi-level Scene Description Network



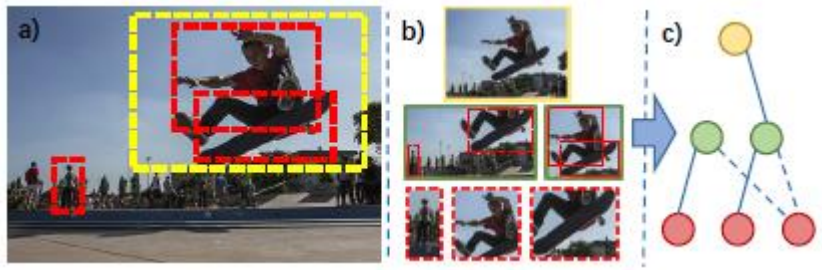
# Multi-level Scene Description Network

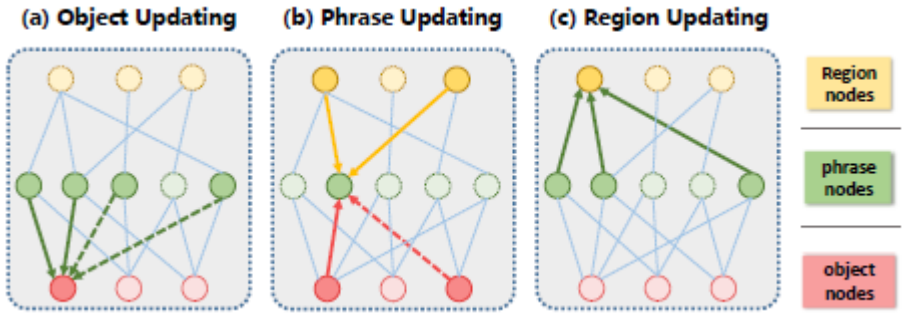


# Multi-level Scene Description Network

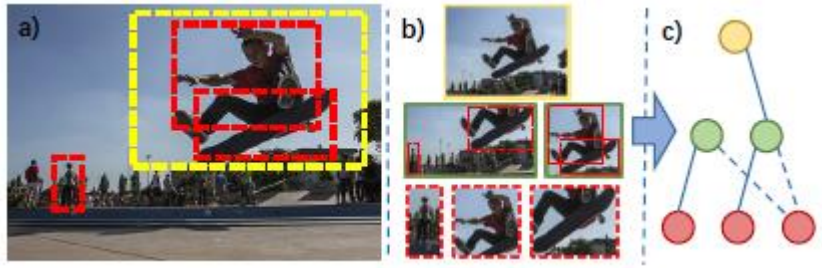


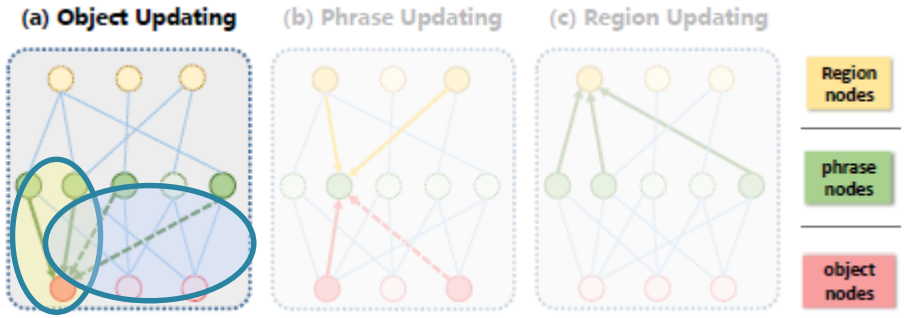
# Multi-level Scene Description Network



# Multi-level Scene Description Network




$$\tilde{x}_i^{(p\to s)} = \frac{1}{\|E_{i,p}\|} \sum_{(i,j)\in E_{s,p}} \sigma_{\langle o,p\rangle}\left(x_i^{(o)}, x_j^{(p)}\right) x_j^{(p)} \quad (1)$$

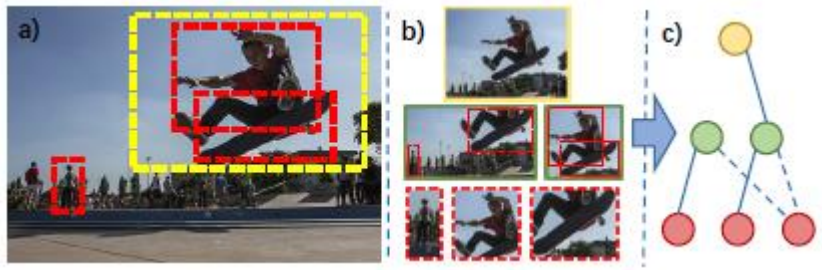
$\tilde{x}_i^{(p\to o)}$ (2)

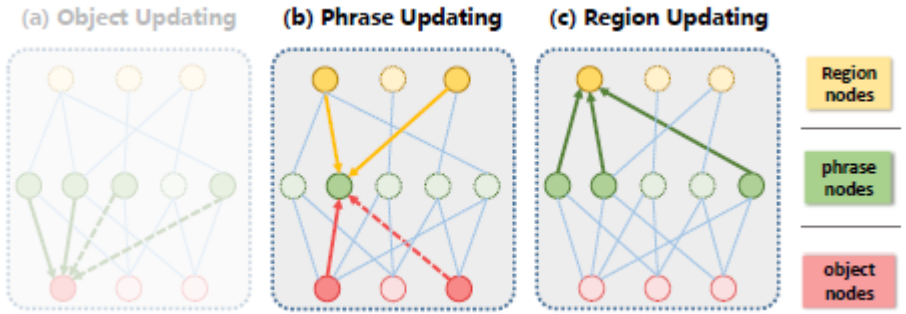
Object feature update위해 (1),(2) 값 사용하여 refining

$$x_{i,t+1}^{(o)} = x_{i,t}^{(o)} + F^{(p\to s)}\left(\tilde{x}_i^{(p\to s)}\right) + F^{(p\to o)}\left(\tilde{x}_i^{(p\to o)}\right) \quad (3)$$

$$F(\cdot) = W \cdot ReLU(\cdot)$$

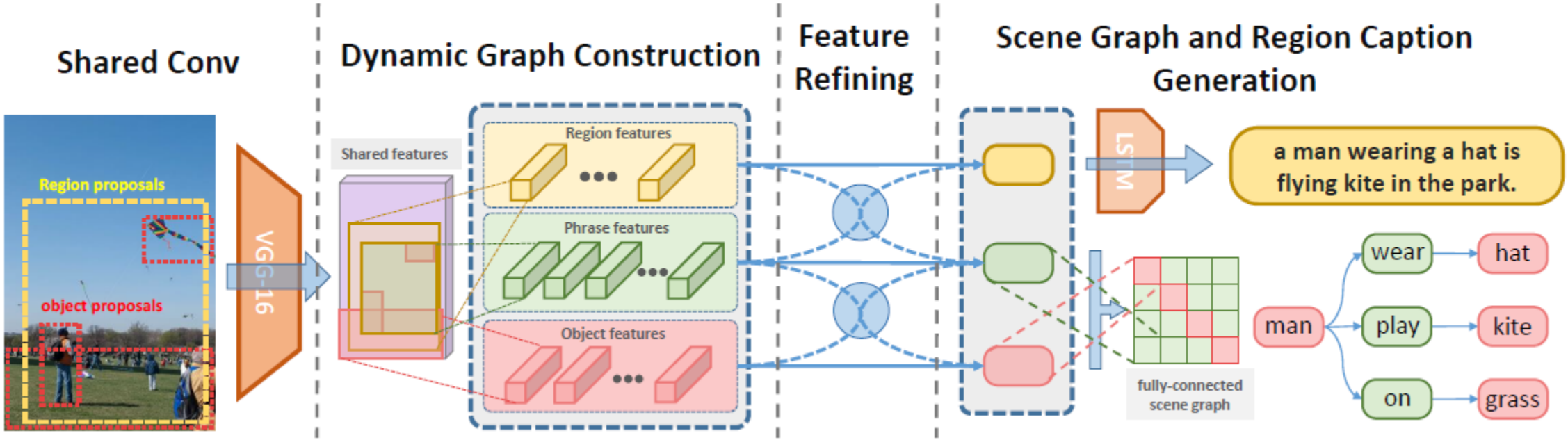
# Multi-level Scene Description Network



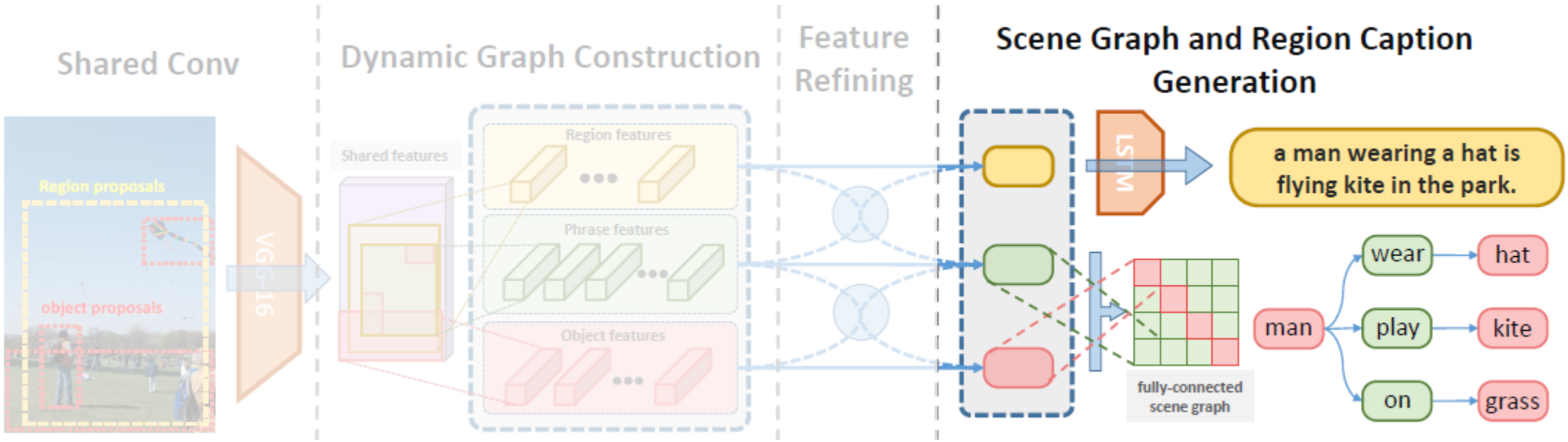




$$
\begin{aligned}
\boldsymbol{x}_{j,t+1}^{(p)} &= \boldsymbol{x}_{j,t}^{(p)} + \boldsymbol{F}^{(s\to p)}\left(\tilde{\boldsymbol{x}}_j^{(s\to p)}\right) \\
&\quad + \boldsymbol{F}^{(o\to p)}\left(\tilde{\boldsymbol{x}}_j^{(o\to p)}\right) + \boldsymbol{F}^{(r\to p)}\left(\tilde{\boldsymbol{x}}_j^{(r\to p)}\right), \quad (4) \\
\boldsymbol{x}_{k,t+1}^{(r)} &= \boldsymbol{x}_{k,t}^{(r)} + \boldsymbol{F}^{(p\to r)}\left(\tilde{\boldsymbol{x}}_k^{(p\to r)}\right),
\end{aligned}
$$

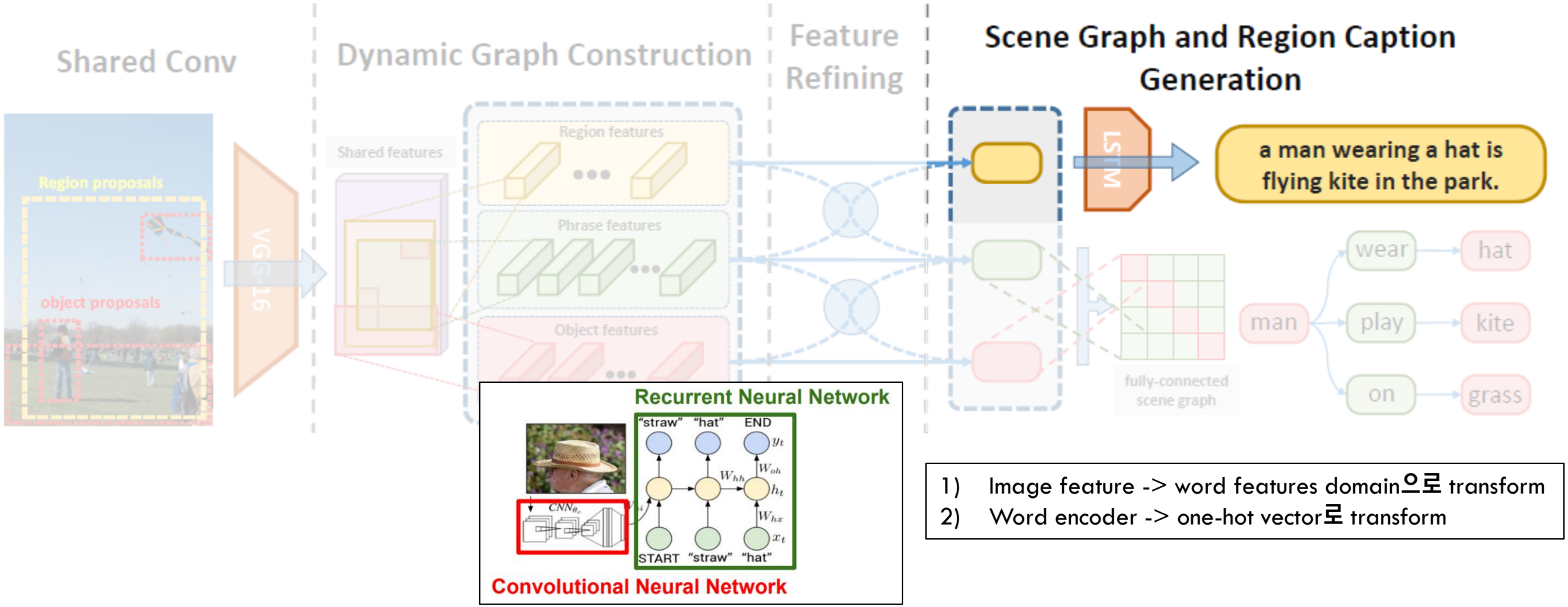
# Multi-level Scene Description Network



# Multi-level Scene Description Network

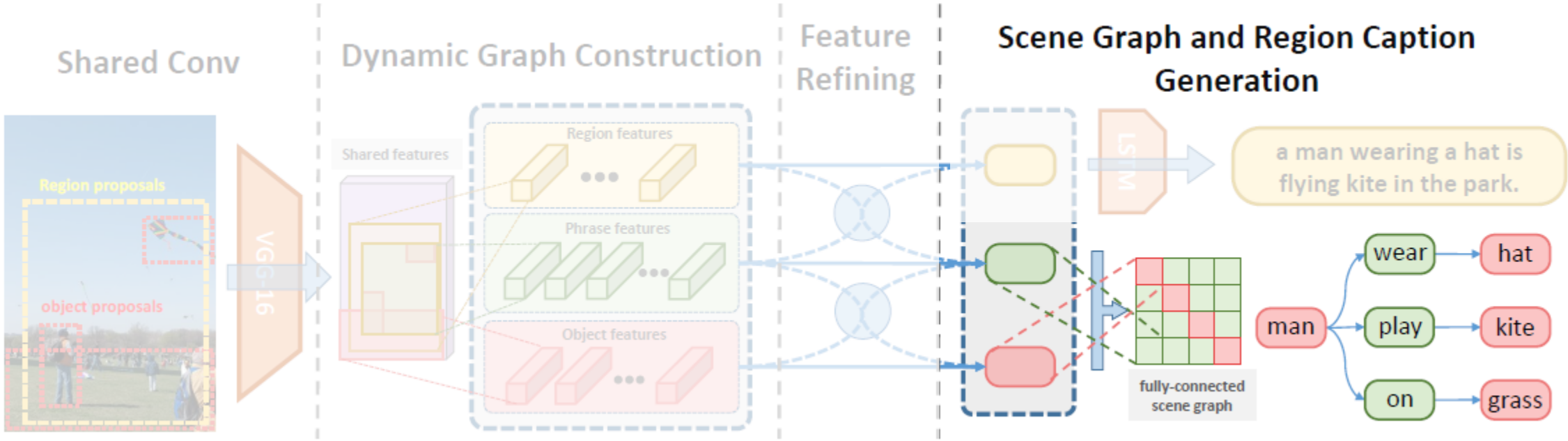


# Multi-level Scene Description Network

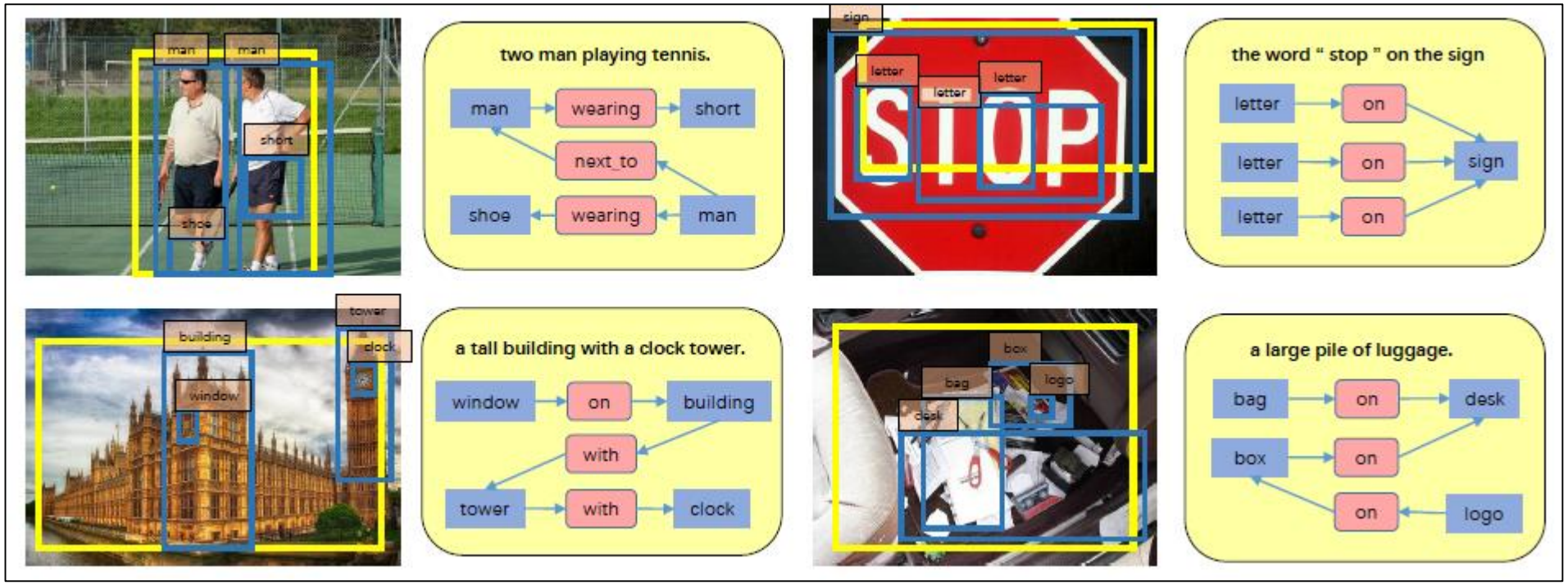



1) Image feature -> word features domain으로 transform
2) Word encoder -> one-hot vector로 transform

# Multi-level Scene Description Network



# The connection is built by proposed dynamic graph generation



# Experiment

From the 95,998 images in the dataset, 25,000 images are sampled as the testing set and the remaining 70,998 images are used as the training set.
(All the experiments are done on the Visual Genome.)

| ID | Message Passing | Cap. branch | Cap. Supervision | FR-iters | PredCls | | PhrCls | | SGGen | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Rec@50 | Rec@100 | Rec@50 | Rec@100 | Rec@50 | Rec@100 |
| 1 | - | - | - | 0 | 49.28 | 52.69 | 7.31 | 10.48 | 2.39 | 3.82 |
| 2 | ✓ | - | - | 1 | 63.12 | 66.41 | 19.30 | 21.82 | 7.73 | 10.51 |
| 3 | ✓ | ✓ | - | 1 | 63.82 | 67.23 | 20.91 | 23.09 | 8.20 | 11.35 |
| 4 | ✓ | ✓ | ✓ | 1 | 66.70 | **71.02** | 23.42 | 25.68 | 10.23 | 13.89 |
| 5 | ✓ | ✓ | ✓ | 2 | **67.03** | 71.01 | **24.22** | **26.50** | **10.72** | **14.22** |
| 6 | ✓ | ✓ | ✓ | 3 | 66.23 | 70.43 | 23.16 | 25.28 | 10.01 | 13.62 |

- Top-K recall(Rec@K)이 main performance metric으로, top-K predictions가 ground truth instances의 부분일 경우
- PredCls : classification of the predicates
- PhrCls : classification of both predicate and object
- SGGen : Object bounding box가 0.5보다 크며 <Subject-Predicate-Object> classification

# Experiment

From the 95,998 images in the dataset, 25,000 images are sampled as the testing set and the remaining 70,998 images are used as the training set.
(All the experiments are done on the Visual Genome.)

| Task | | LP [31] | ISGG [40] | Ours |
|---|---|---|---|---|
| PredCls | R@50 | 26.67 | 58.17 | **67.03** |
| | R@100 | 33.32 | 62.74 | **71.01** |
| PhrCls | R@50 | 10.11 | 18.77 | **24.34** |
| | R@100 | 12.64 | 20.23 | **26.50** |
| SGGen | R@50 | 0.08 | 7.09 | **10.72** |
| | R@100 | 0.14 | 9.91 | **14.22** |

LP(Language Prior) : Object detection 후, 위치 관계를 사용하여 word embedding 된 predicate categories 결정
ISGG(Iterative Scene Graph Generation) : GRU-based feature refining scheme.(caption branch 부분이 없음)

# Conclusion

- Scene Graph Generation에 있어서 Object, Phrase 외에 Caption 정보를 사용하여 학습하면 좋은 성능을 냄
- 아쉬웠던 부분

 - Feature Refining 부분에서 gate templates이 명확하지 않아서 edge 선택 기준이 모호했던 점