موضوع

YOLO v12

قهرمان تشخیص اشیا :)

YOLOv12: Attention-Centric Real-Time Object Detectors

Yunjie Tian
University at Buffalo
yunjieti@buffalo.edu

Qixiang Ye
University of Chinese Academy of Sciences
qxye@ucas.ac.cn

David Doermann
University at Buffalo
doermann@buffalo.edu

موضوع

YOLO v12

قهرمان تشخیص اشیا :)

YOLOv12: Attention-Centric Real-Time Object Detectors

Yunjie Tian
University at Buffalo
yunjieti@buffalo.edu

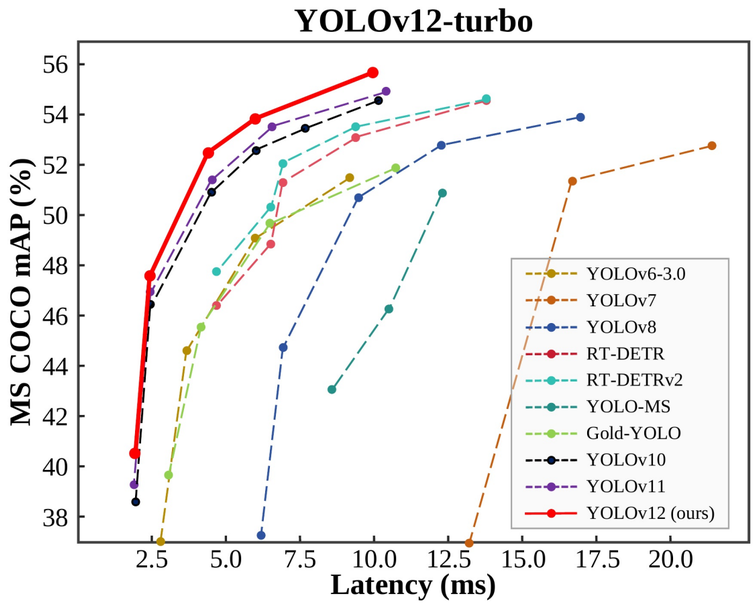
Qixiang Ye
University of Chinese Academy of Sciences
qxye@ucas.ac.cn

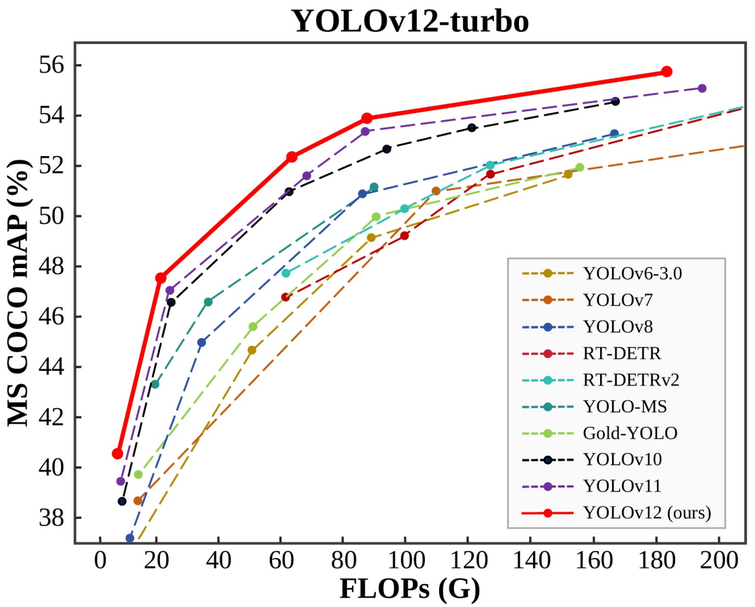
David Doermann
University at Buffalo
doermann@buffalo.edu

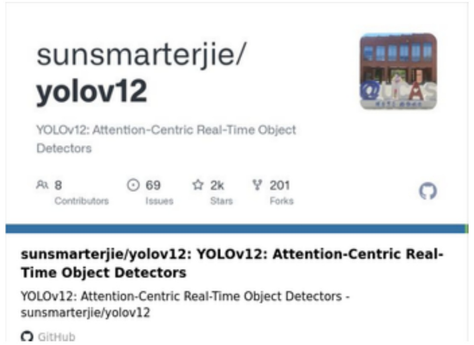
کاربردها و موارد استفاده YOLOV12

- شناسایی اشیاء (Object Detection)
- بخش بندی نمونه ها (instance segmentation)
- دسته بندی تصاویر (Image Classification)
- تخمین وضعیت بدن (Pose Estimation)
- تشخیص اشیا با باکس های مرزی مورب ((Oriented Object Detection (OBB

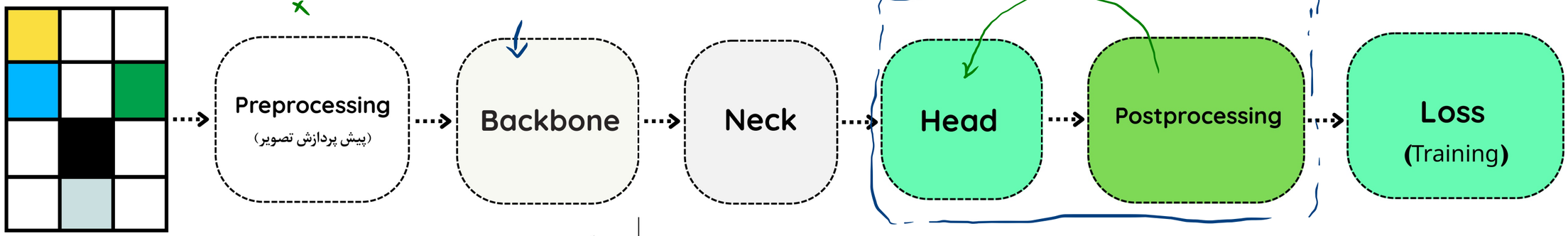
چرا؟


YOLOv12-turbo
MS COCO mAP (%)
Latency (ms)
YOLOv6-3.0
YOLOv7
YOLOv8
RT-DETR
RT-DETRv2
YOLO-MS
Gold-YOLO
YOLOv10
YOLOv11
YOLOv12 (ours)


YOLOv12-turbo
MS COCO mAP (%)
FLOPs (G)
YOLOv6-3.0
YOLOv7
YOLOv8
RT-DETR
RT-DETRv2
YOLO-MS
Gold-YOLO
YOLOv10
YOLOv11
YOLOv12 (ours)


sunsmarterjie/
yolov12
YOLOv12: Attention-Centric Real-Time Object Detectors
8 Contributors
69 Issues
2k Stars
201 Forks
sunsmarterjie/yolov12: YOLOv12: Attention-Centric Real-Time Object Detectors
YOLOv12: Attention-Centric Real-Time Object Detectors - sunsmarterjie/yolov12
GitHub

معماری کلی YOLOv12



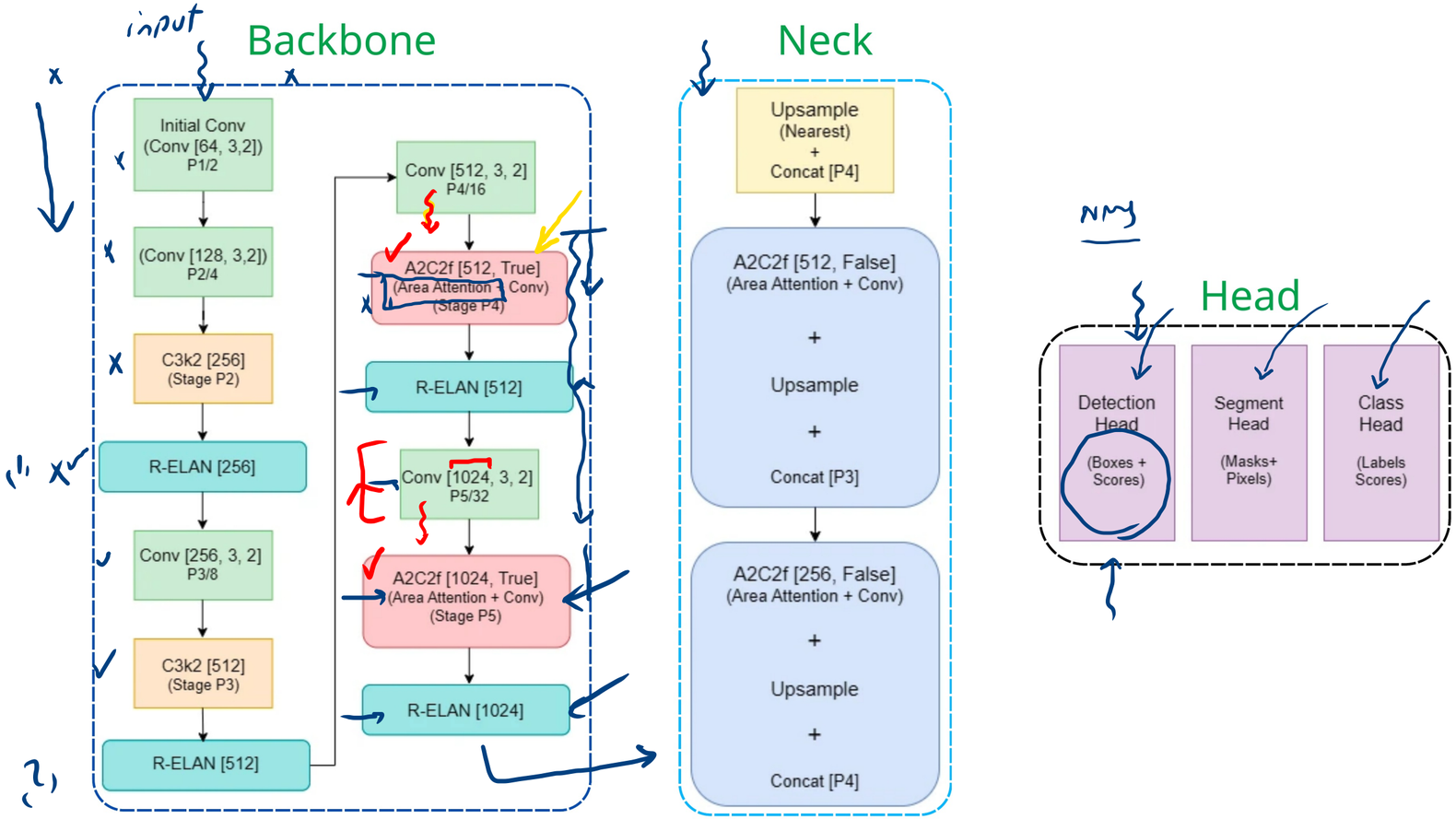
- ویژگی‌های مهم از تصویر استخراج می شود!
- ترکیب ویژگی ها و کمک به مدل که اشیاء در سایزهای مختلف را راحت‌تر ببیند.
- مختصات جعبه + کلاس شیء را پیش‌بینی می‌کند.

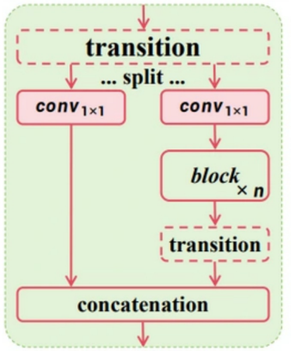
Non-Maximum Suppression (NMS) برای حذف باکس های تکراری.

confidence threshold : برای حذف باکس های غیرقابل اعتماد

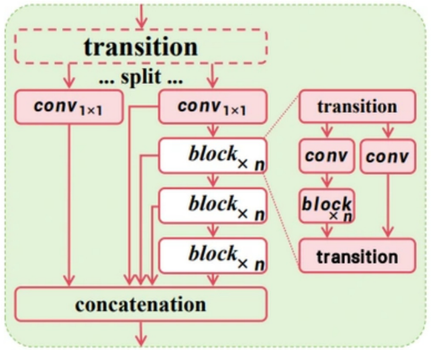
- پیش‌بینی‌ها مقایسه می‌شود با برچسب‌های واقعی (Ground Truth)
- Loss محاسبه می‌شود (مثل Box Loss، Classification Loss، Mask Loss، Keypoint Loss).
- از این Loss برای آپدیت وزن‌های مدل استفاده می‌شود (Backpropagation + Optimizer).

معماری جزیی YOLOv12

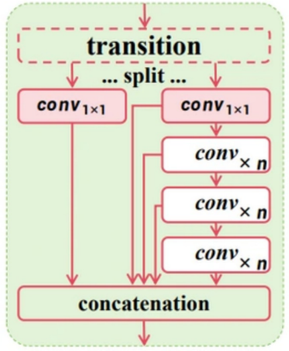




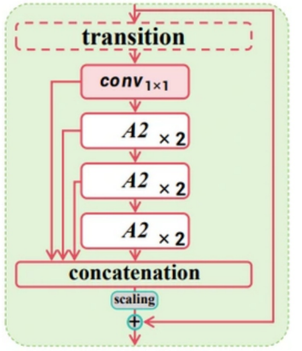
(a) CSPNet



(c) C_3K_2

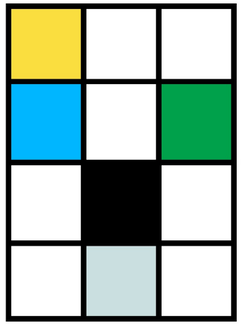


(b) ELAN



(d) R-ELAN

مرحله ۱ : پیش پردازش



Preprocessing

(پیش پردازش تصویر)

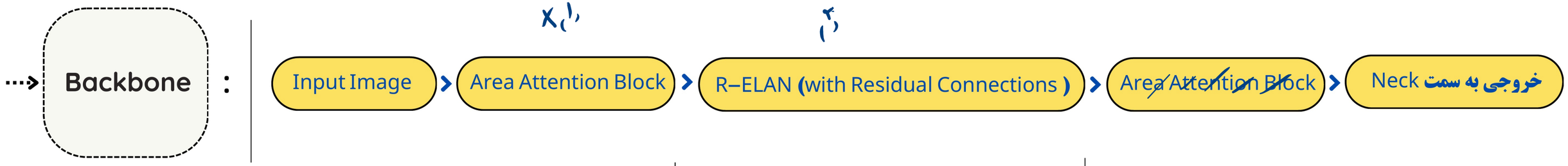
:

- **Resize : 640x640, ...**
- **Normalize : (Image/255)**
- **Data Augmentation Pipeline**

  Flipping, Cropping, Scaling, Mosaic Augmentation , ...

مرحله۲ : Backbone

در این مرحله ویژگی‌های مهم از تصویر استخراج می شود! (از ویژگی‌های ساده مثل لبه‌ها، تا ویژگی‌های پیچیده مثل شکل کلی اشیاء)



- قرارگیری در مراحل ابتدایی و میانی Backbone
- توجه به نواحی به جای توجه به تک تک پیکسل ها
- توجه بیشتر به مناطقی که احتمال وجود شی بیشتر است.
- سرعت بالا همراه با دقت بالا
- خیلی سبک تر از self attention در ترنسفورمر

- باعث می شود ویژگی‌های دقیق‌تر و غنی‌تر استخراج شوند.
- باعث می شوند که مدل بتوند اشیاء کوچک، متوسط و بزرگ را با هم خوب تشخیص بدهد.

مرحله۱ از بخش Backbone : قسمت Area Attention Block

توجه بیشتر به مناطقی که احتمال وجود شی بیشتر است.

Area Attention Block

تمرکز روی نواحی به جای توجه تمرکز روی پیکسل ها
(برخلاف self attention)

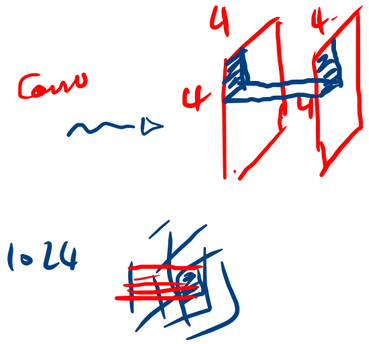
سرعت بیشتر نسبت به self attention
کاهش هزینه محاسباتی

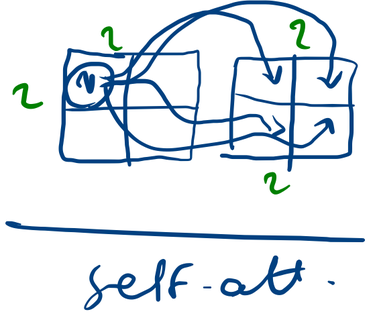
Area Attention

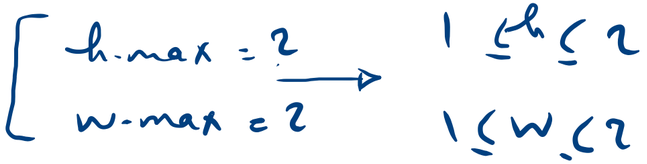
Yang Li [1] **Lukasz Kaiser** [1] **Samy Bengio** [1] **Si Si** [1]

[1]Google Research, Mountain View, CA, USA. Correspondence to: Yang Li <liyang@google.com>.

Area Attention Block – 1

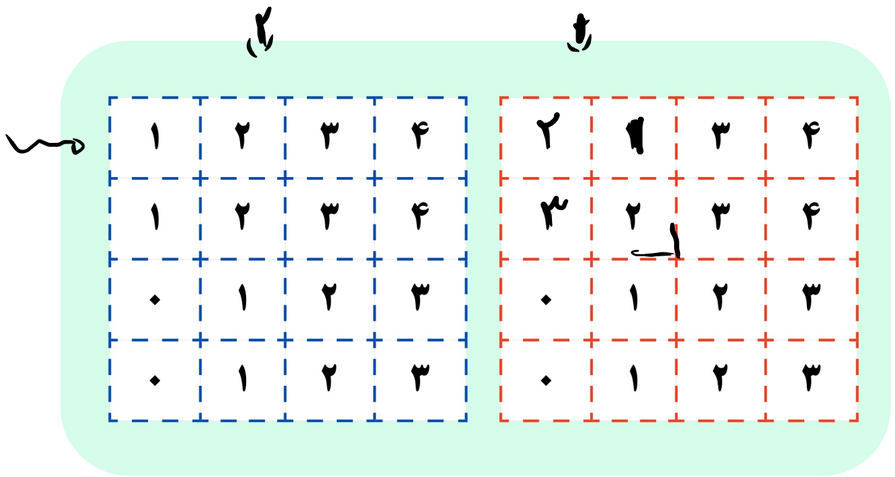




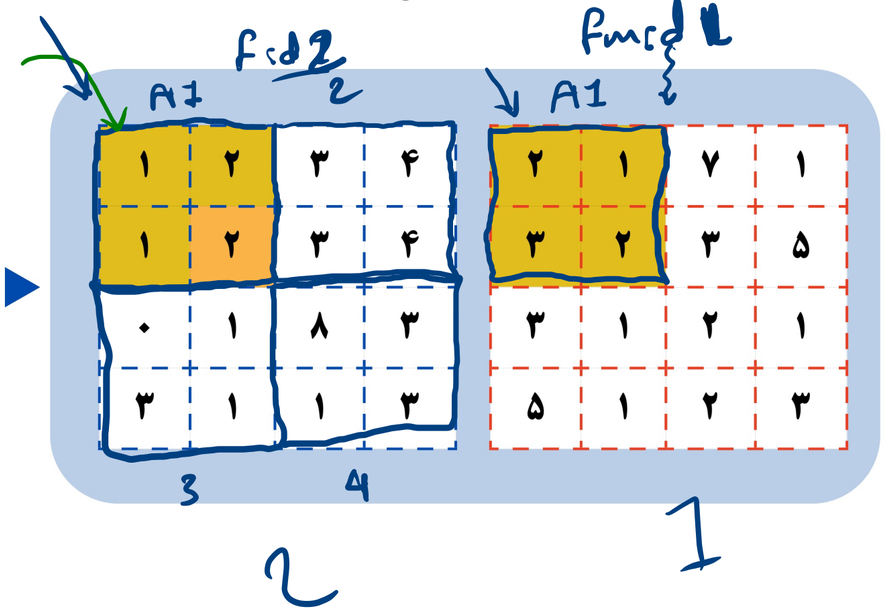


تصویر (فیچر مپ) ورودی

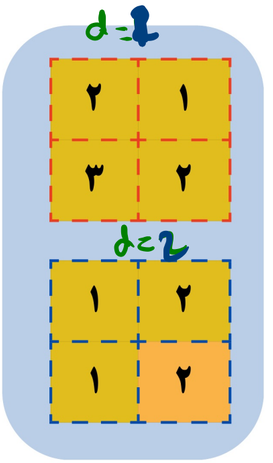
(shape 4×4×2)



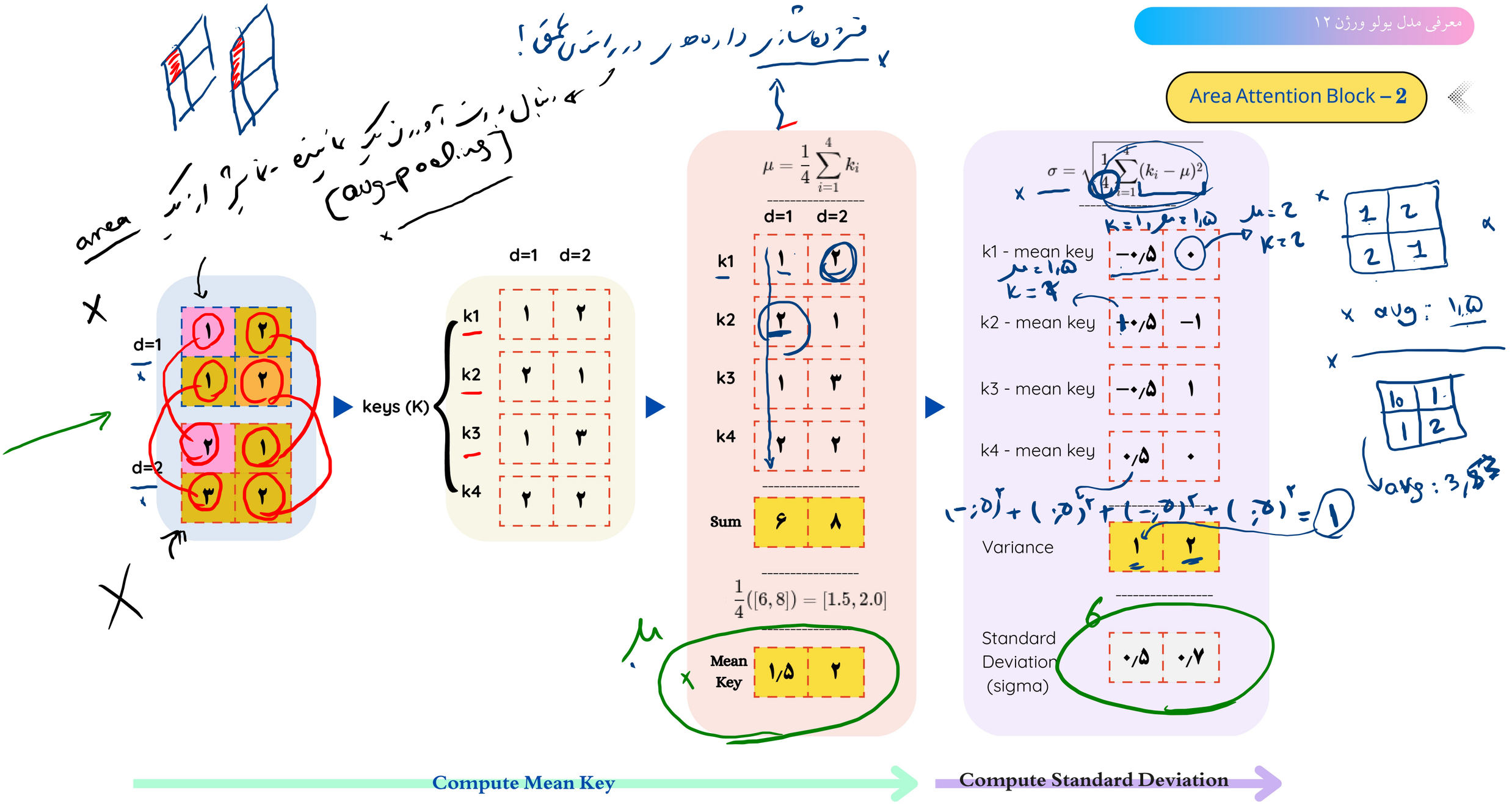
2×2 region (area)



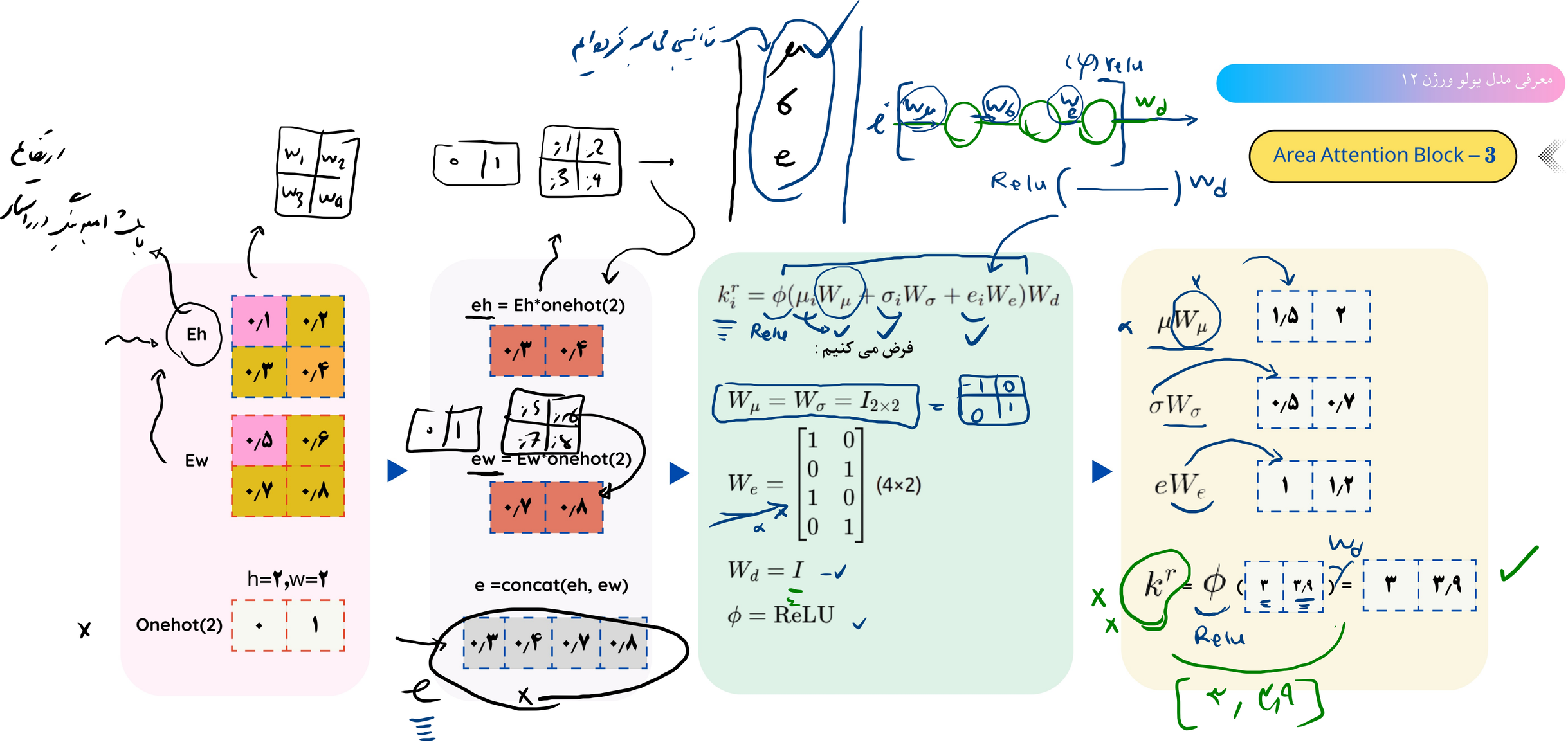
area (0, 0)



Area Attention Block – 2



Area Attention Block – 3



Area Shape Embedding

Combine into Final Area Key

Area Attention Block – 4

$$\frac{e^{i}}{\sum e^{i}} = \frac{e^{6.9}}{e^{6.9} + e^{8} + e^{14} + e^{7.7}}$$

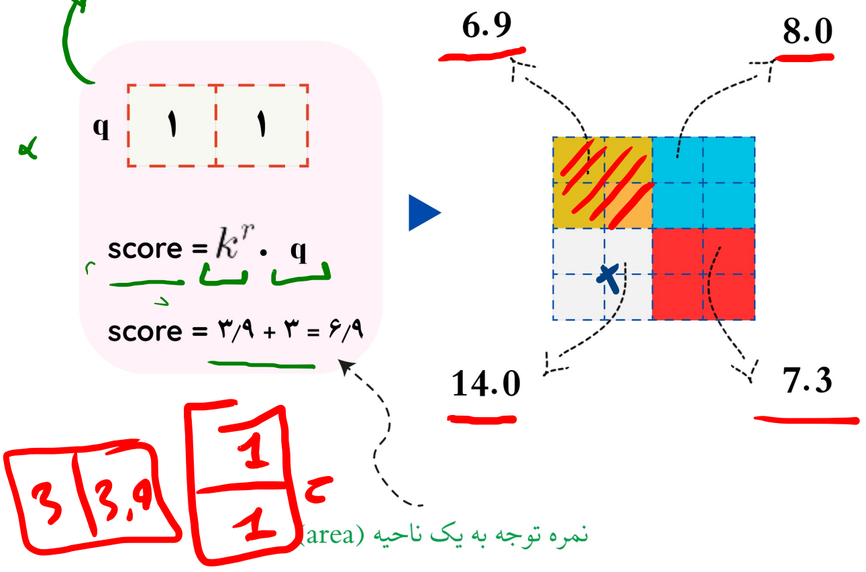
W

parameter.

trainable

وزن

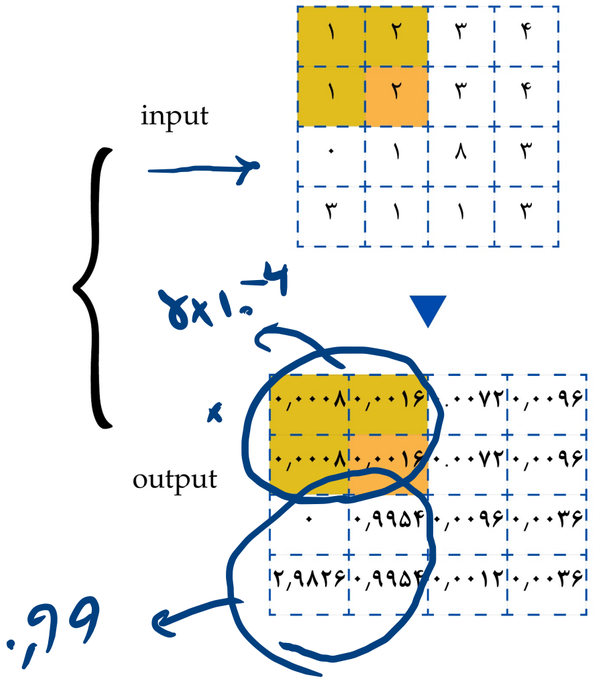
فرض: محاسبه چند ناحیه!



Scores = [6.9, 8.0, 14.0, 7.3]
attention weights = Softmax(Scores)

attention weights =
[0.00082, 0.00247, 0.99549, 0.00123]

۰/۰۰۰۸	*	top left
۰/۰۰۲۴	*	top right
۰/۹۹۵۴	*	bottom left
۰/۰۰۱۲	*	bottom right

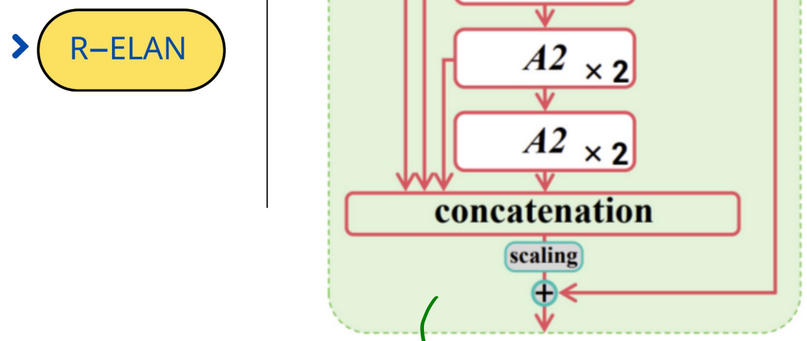


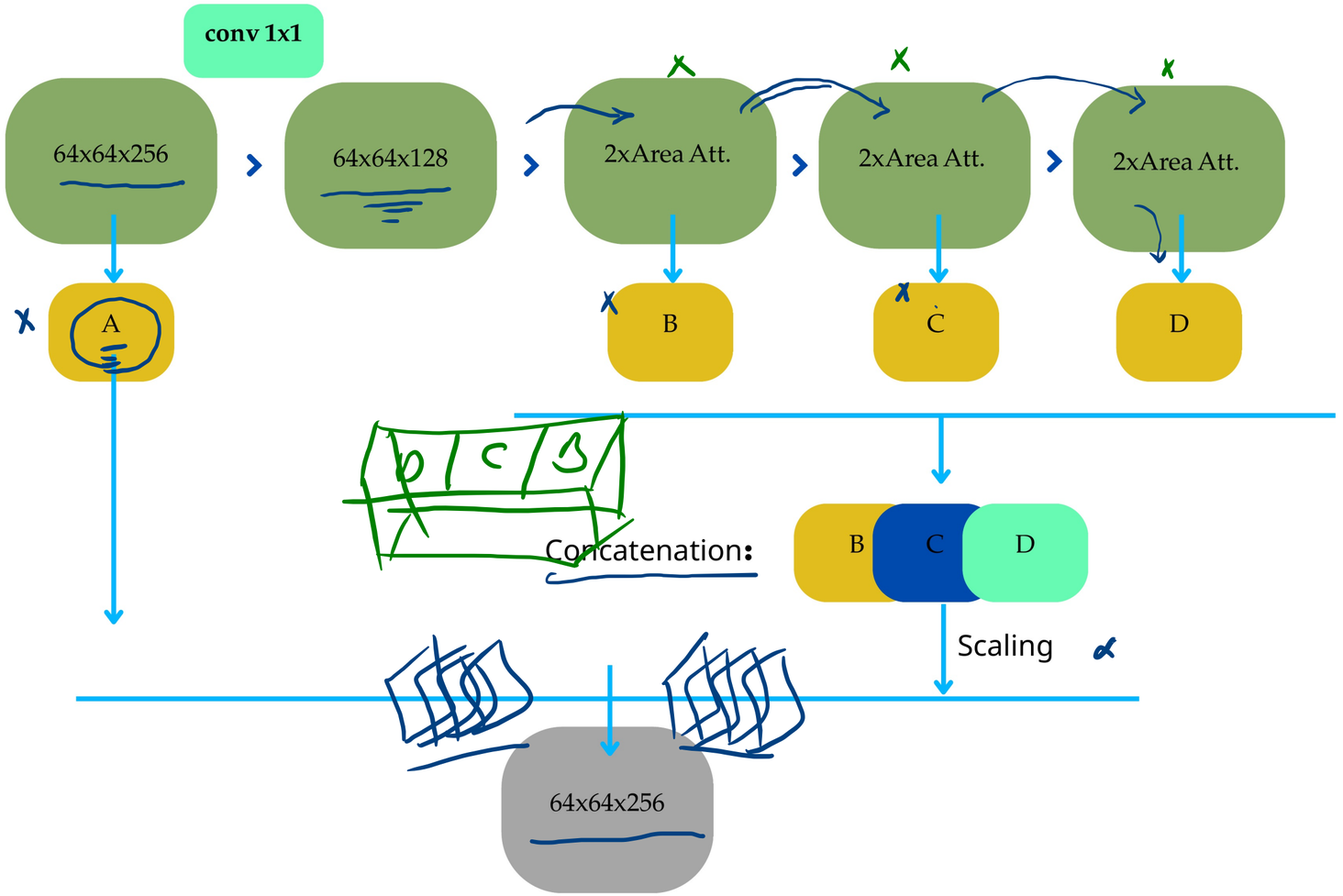
Query & Dot Product

Final Attention Output

Residual–Enhanced Efficient Layer Aggregation Network

R-ELAN





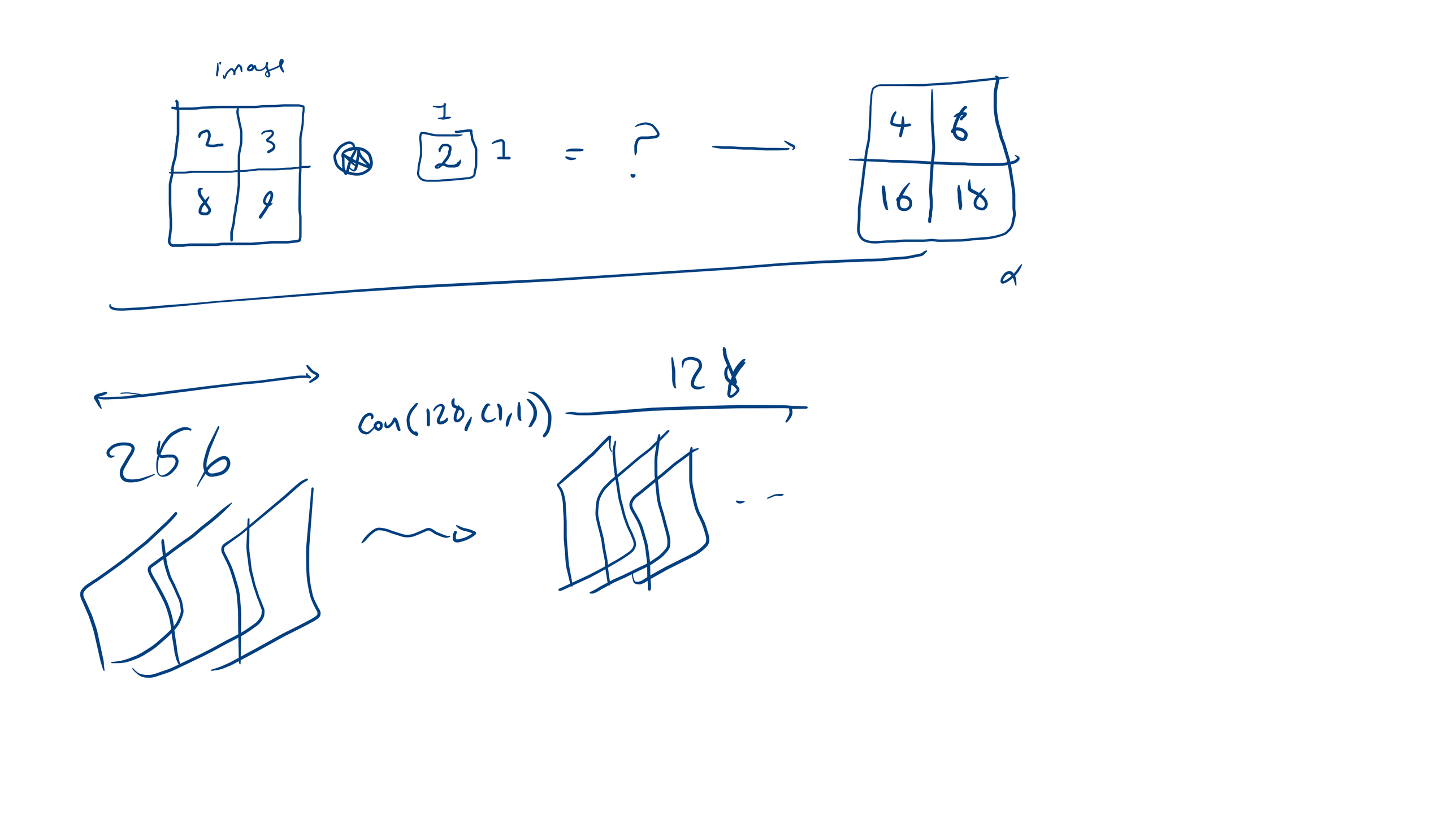
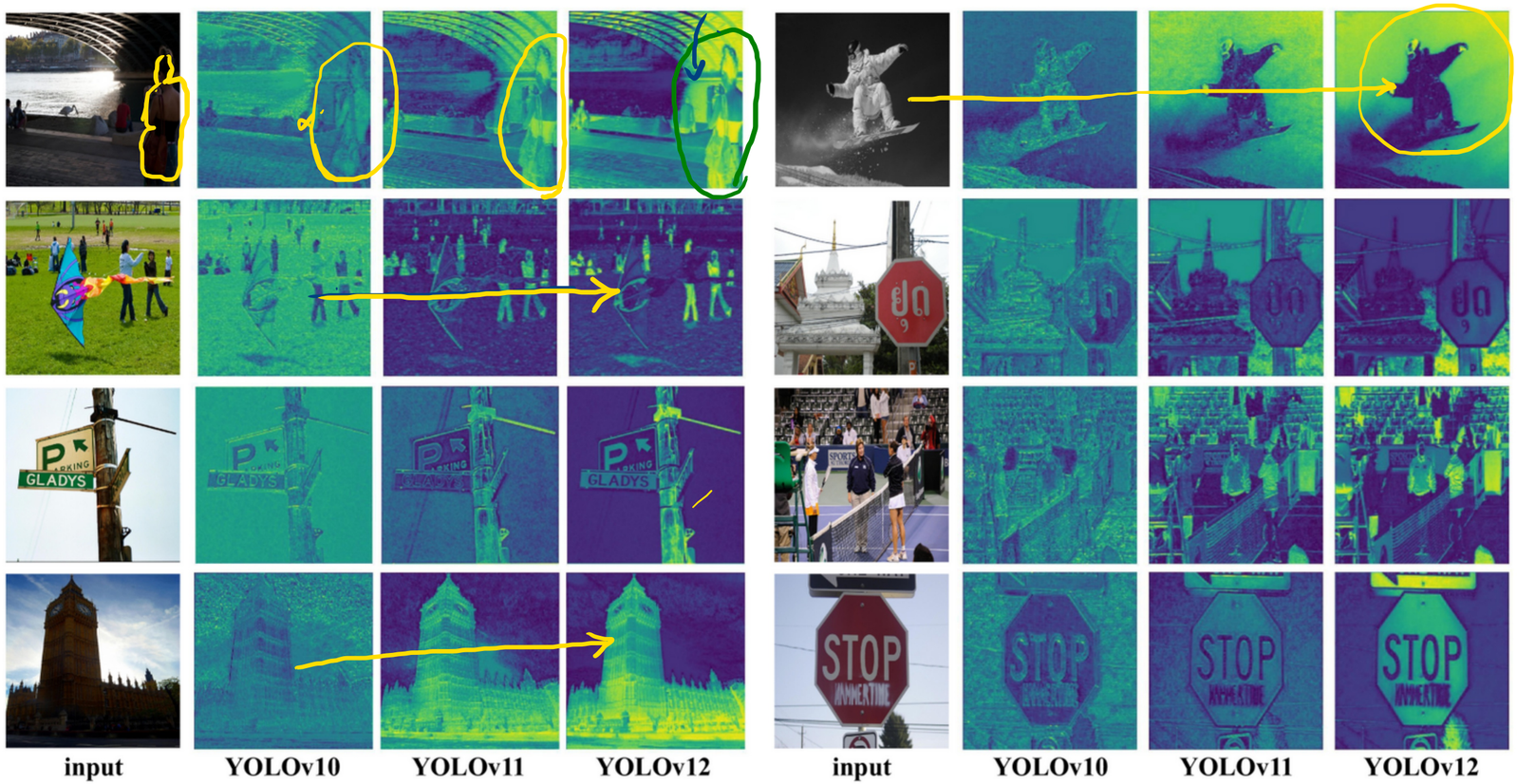


image
2
3
8
9
1
2
1
=
?
4
6
16
18
256
Con(128, (1,1))
128

قسمت توجه مربوط به BACKBONE

خروجی یولو ۱۲ با AREA ATTENTION




The End