Aligning Pretraining for Detection via Object-Level Contrastive Learning

Microsoft Research Asia

韩坤洋

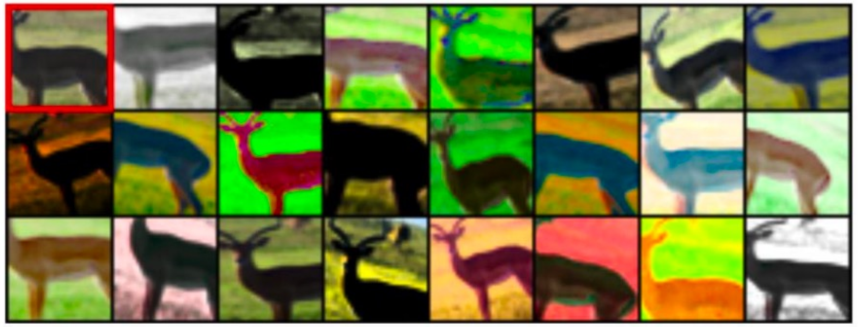
Self-supervised pretraining

- Segmentation supervised
  - Image -> Logits mask prediction -> CELoss <- GT mask (Human annotate)
  - Image (-> Semantic information) -> Logits mask prediction

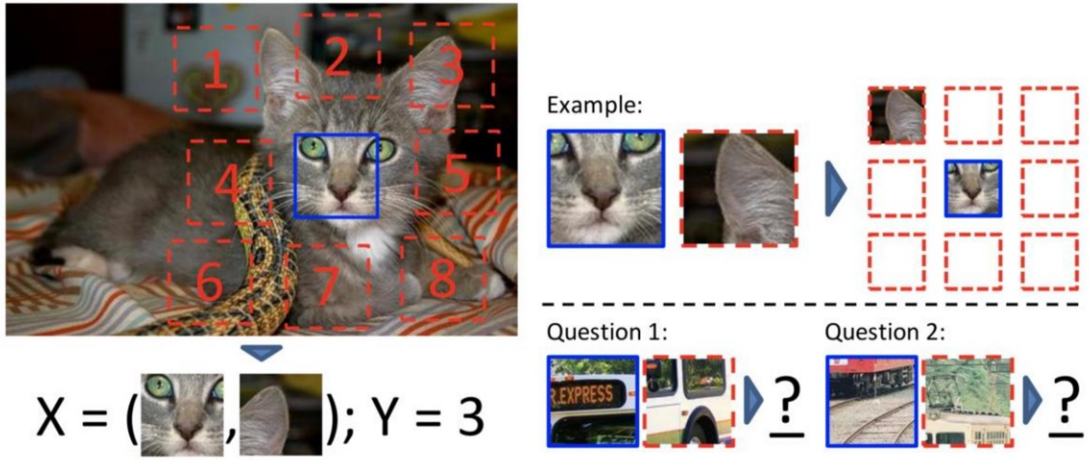
- Segmentation self-supervised
  - Image (-> Semantic info) -> Pretext pred -> Loss <- Pretext GT (Generated)
  - **Image (-> Semantic info)** -> Logits mask prediction -> CELoss <- GT mask

Self-Supervised Pretext

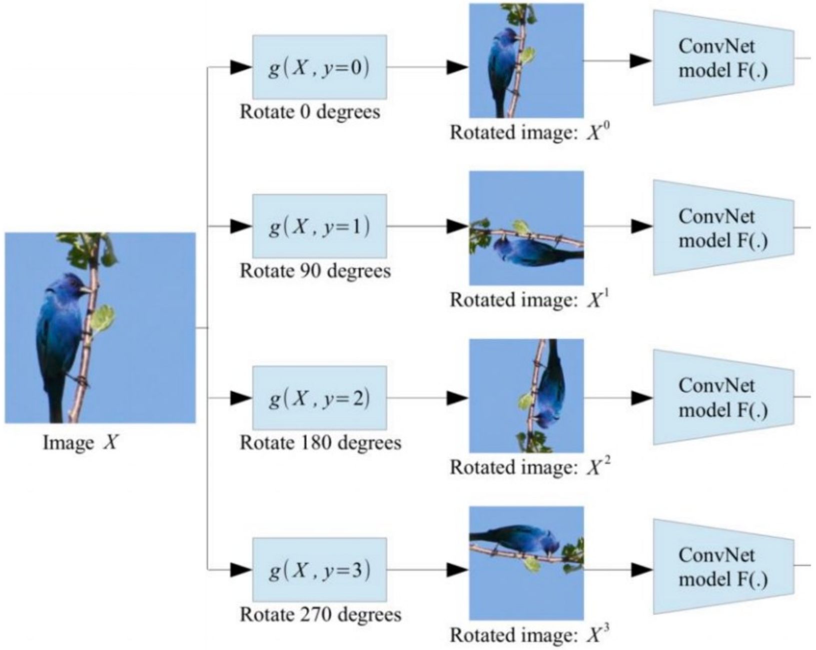
- Distortion



- Patches



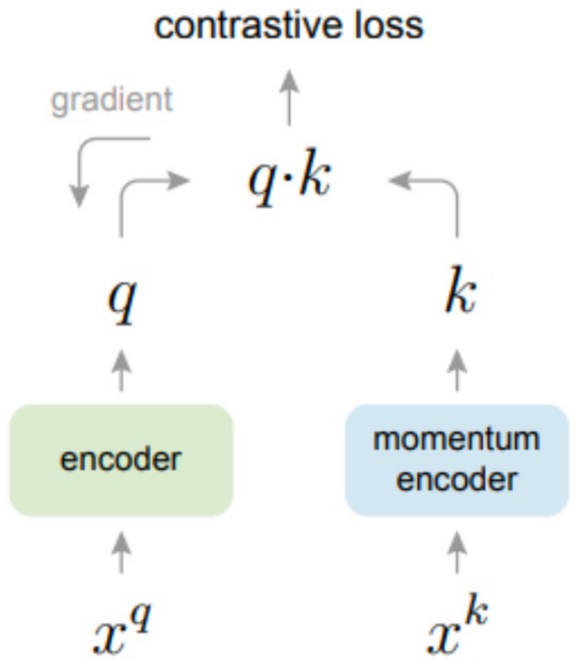
- Rotation



MoCo - Dictionary Look-up

- Keys in the dictionaries
  - Sample from data, images or patches
  - Represented by encoder network
- Encoded 'query'
  - Similar to its matching 'key'
  - Dissimilar to others
- Loss

$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i / \tau)}$$

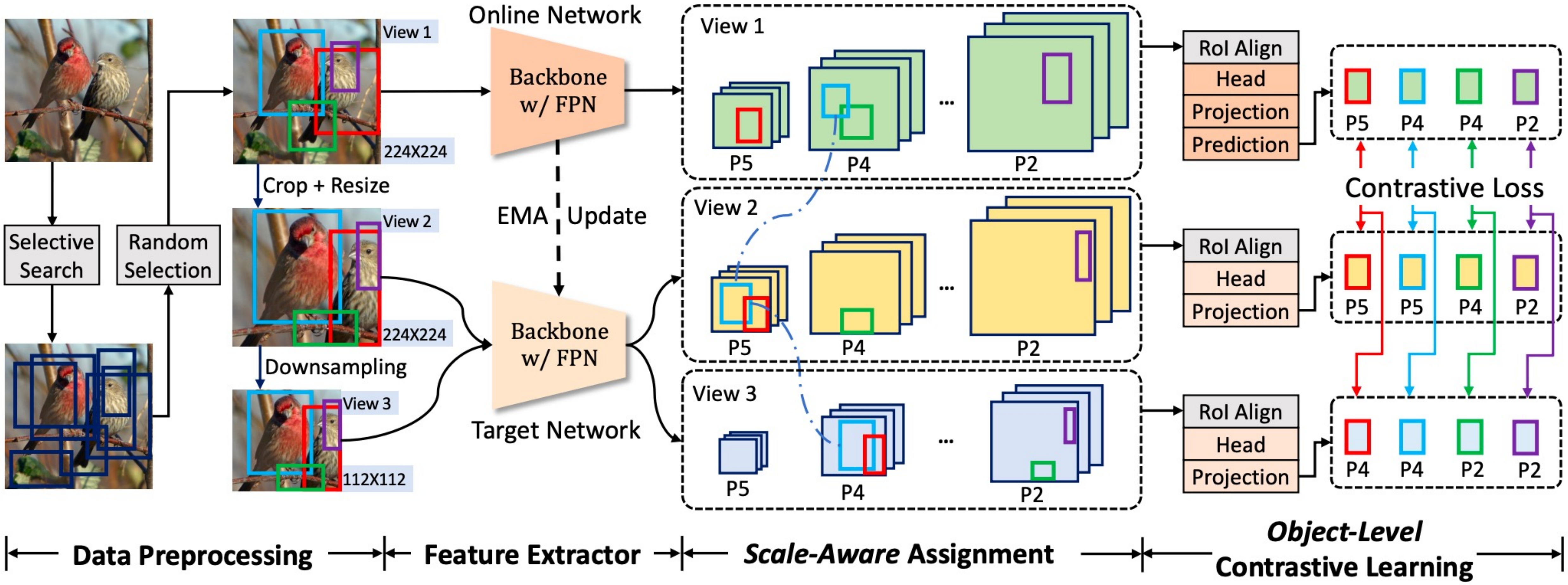


Motivation

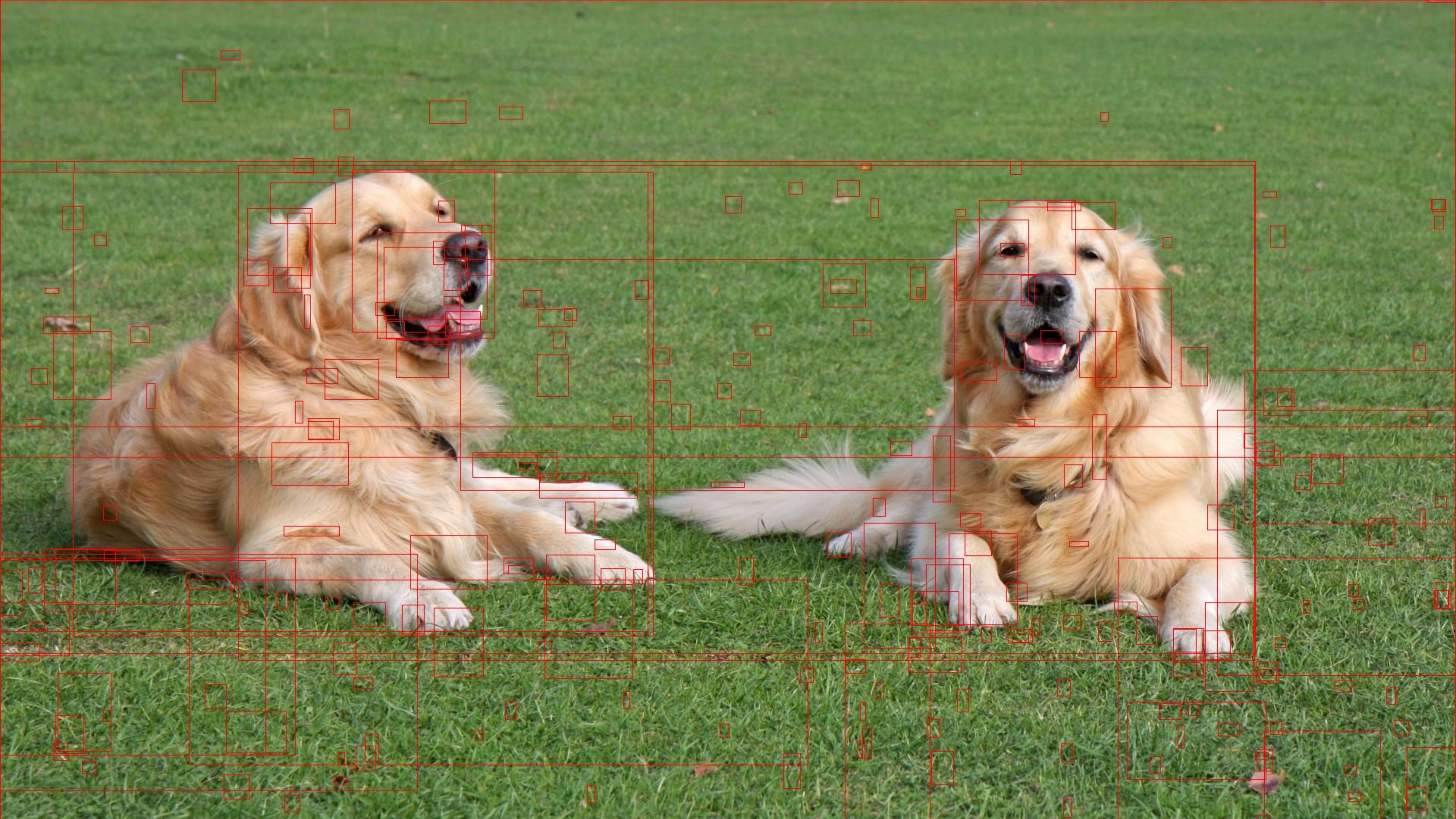
- Image-level are sub-optimal for dense prediction tasks
  - object detection, semantic segmentation
- May overfit to holistic representations

- Goal: develop self-supervised pretraining aligned to object detection.
  - Translation and scale invariance

Arch

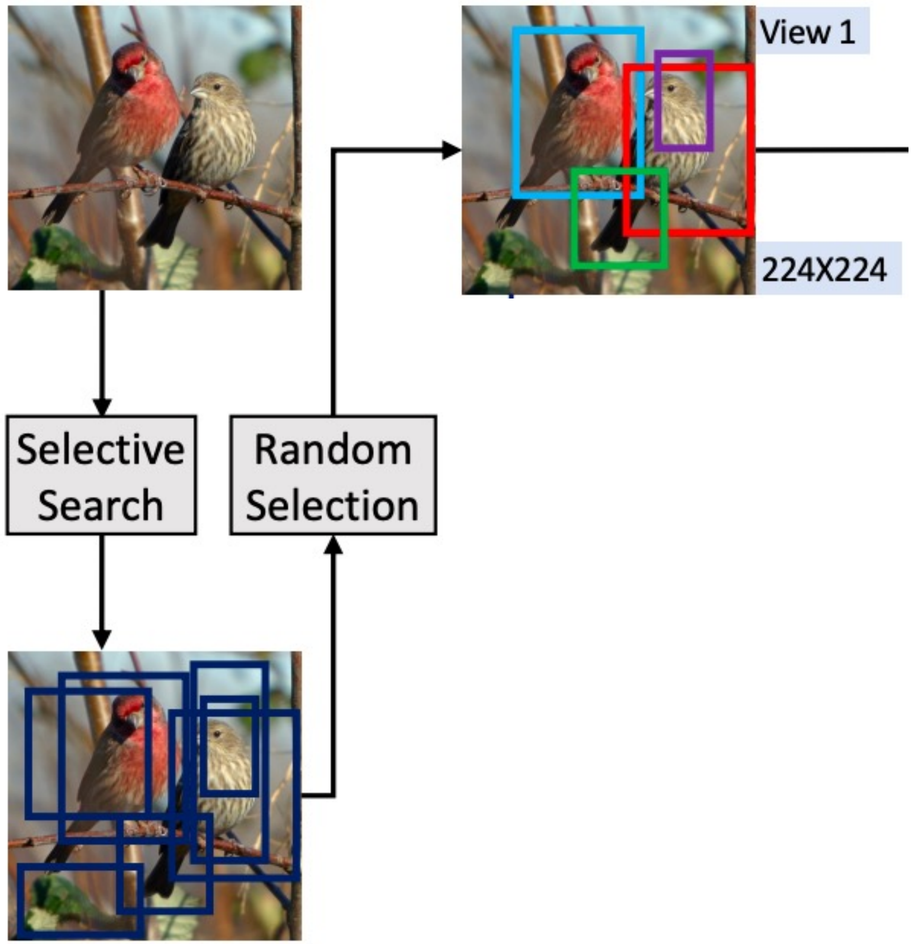


View 1
224X224
Crop + Resize
View 2
224X224
Downsampling
View 3
112X112
Selective Search
Random Selection
Online Network
Backbone w/ FPN
EMA Update
Backbone w/ FPN
Target Network
View 1
P5
P4
P2
View 2
P5
P4
P2
View 3
P5
P4
P2
RoI Align
Head
Projection
Prediction
P5
P4
P4
P2
Contrastive Loss
RoI Align
Head
Projection
P5
P5
P4
P2
RoI Align
Head
Projection
P4
P4
P2
P2
Data Preprocessing
Feature Extractor
Scale-Aware Assignment
Object-Level Contrastive Learning



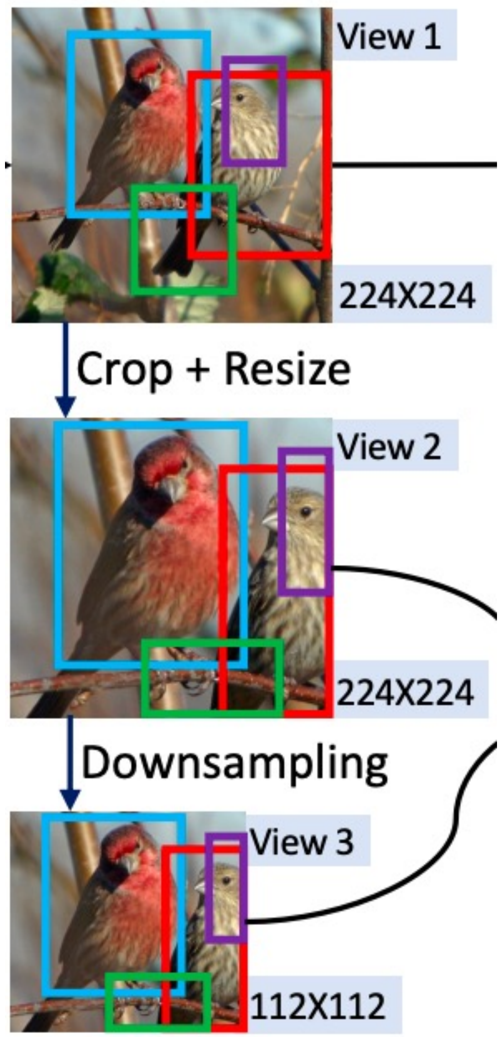
Object proposal

- Selective search,
  - unsupervised algorithm
- 1) $1/3 \leq w/h \leq 3$; 2) $0.3 \leq \sqrt{wh}/\sqrt{WH} \leq 0.8$,
- Random select K proposals



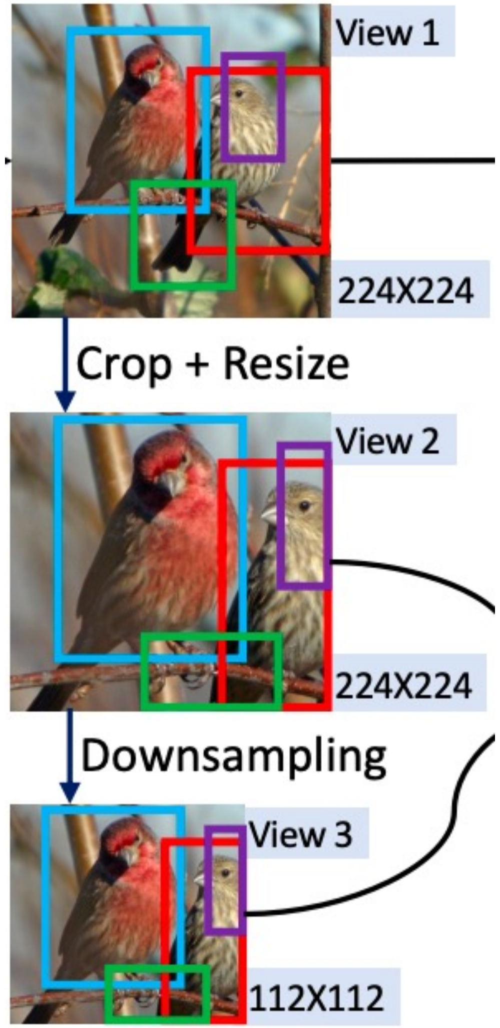
View Construction

- V1
  - Resize to 224 x 224
- V2
  - Random crop, scale [0.5, 1] on V1
  - Resize to 224 x 224
  - Proposals outside of V2 are dropped
- V3
  - Resize to 112 x 112, from V2



View Construction

- V1, V2, V3 randomly and independently augmented
  - Horizontal flip
  - Color distortion
    - Brightness, contrast, saturation, hue adjustments
  - Optional grayscale conversion
  - Gaussian blur
  - Solarization



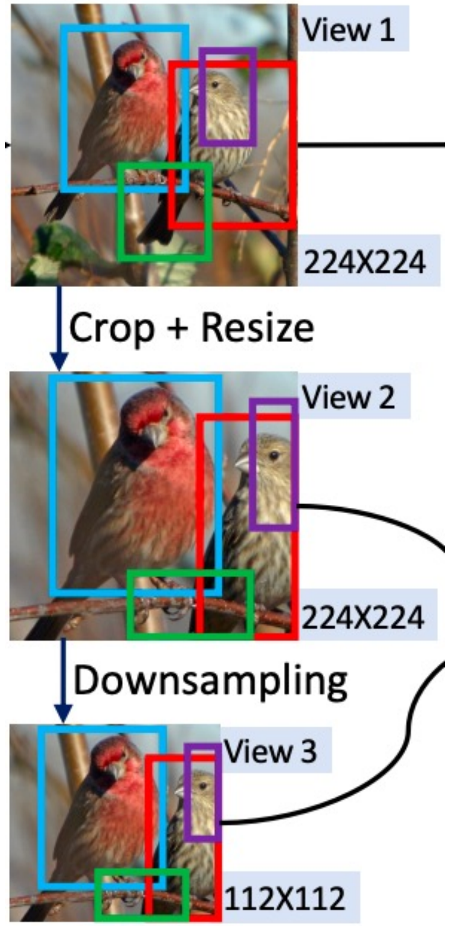
Box Jitter

- To encourage variance of scales and locations
- With a probability of 0.5

$$b = \{x, y, w, h\}$$

$$1)\ \hat{x} = x + r \cdot w;\ 2)\ \hat{y} = y + r \cdot h;$$
$$3)\ \hat{w} = w + r \cdot w;\ 4)\ \hat{h} = h + r \cdot h,$$

$$r \in [-0.1, 0.1]$$



Aligning Architecture to Object Detection

- Two neural network
  - Same arch
  - Different weight

$$\theta_{\mathrm{k}} \leftarrow m\theta_{\mathrm{k}} + (1-m)\theta_{\mathrm{q}}.$$

- FPN

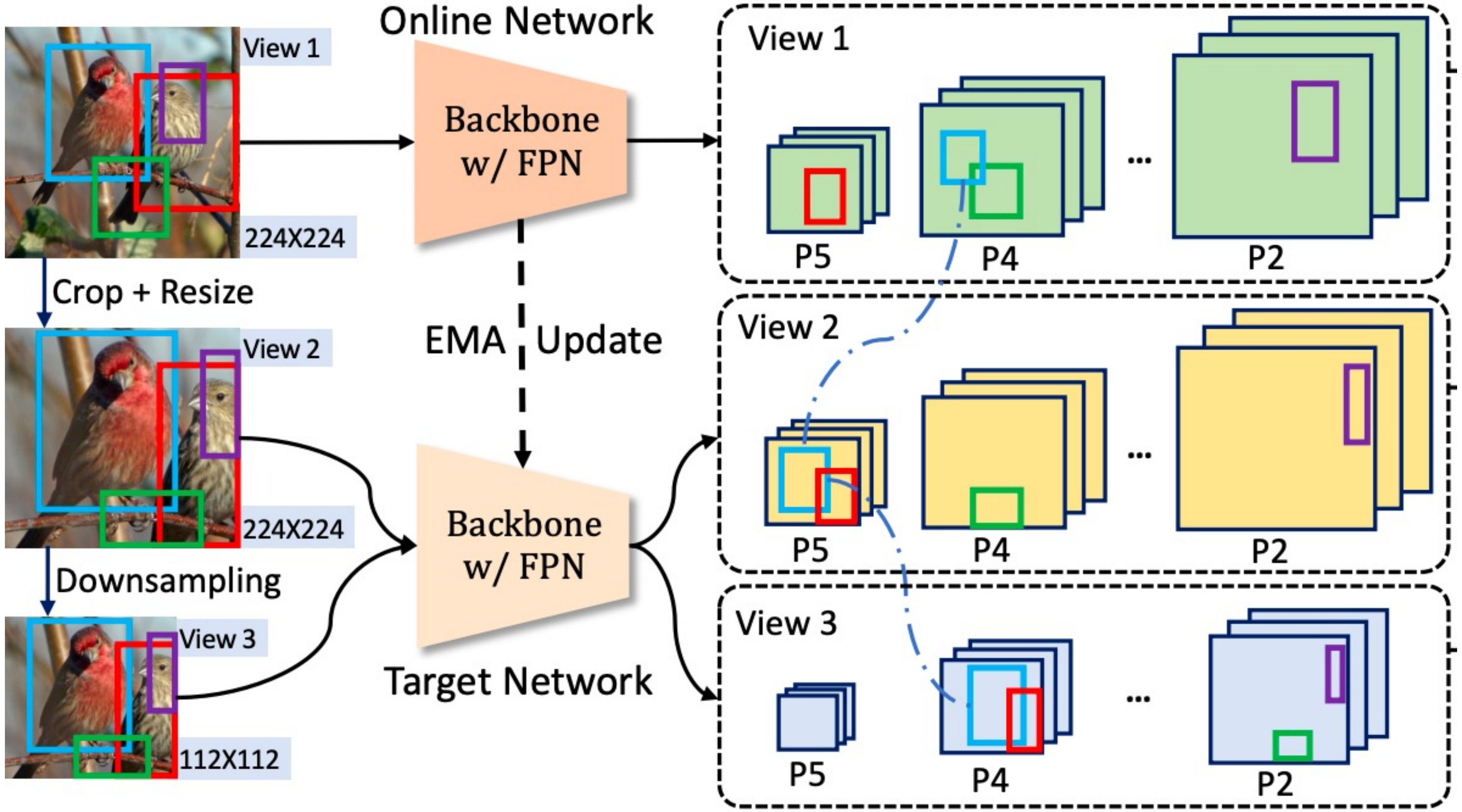
$\{P_2, P_3, P_4, P_5\}$ with a stride of $\{4, 8, 16, 32\}$.

- R-CNN head

$$h = f^H(\text{RoIAlign}(f^I(V), b)).$$

h, object level representation

image-level feature extractor f^I



Aligning Architecture to Object Detection

- Projection

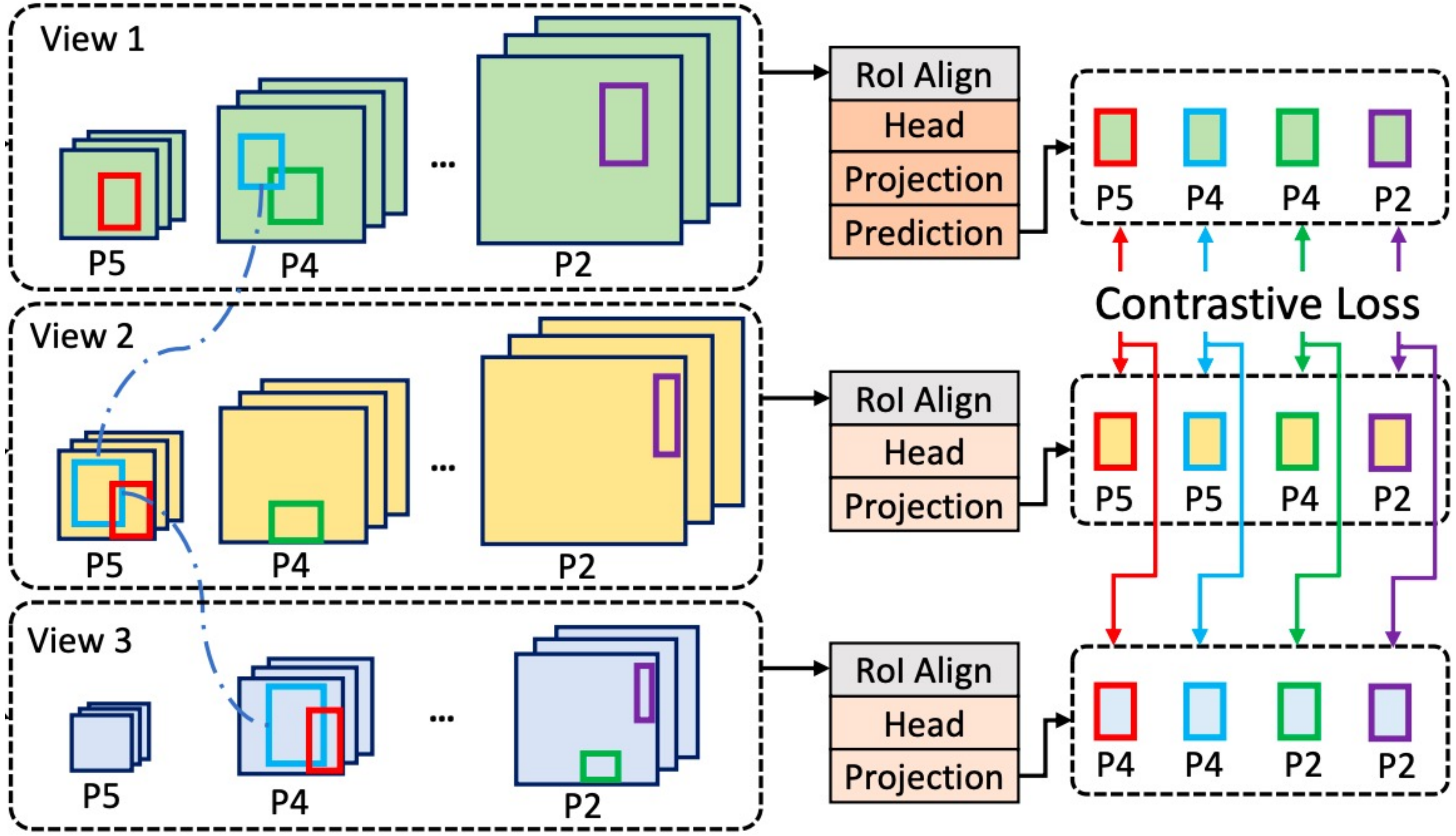
$$v_i = q_\theta(g_\theta(h_i)), \quad v'_i = g_\xi(h'_i), \quad v''_i = g_\xi(h''_i).$$

- Loss

$$\mathcal{L}_i = -2 \cdot \frac{\langle v_i, v'_i \rangle}{\|v_i\|_2 \cdot \|v'_i\|_2} - 2 \cdot \frac{\langle v_i, v''_i \rangle}{\|v_i\|_2 \cdot \|v''_i\|_2}.$$

- Symmetry
  - V2 V3 feed to online network
  - V1 feed to target network

$$\mathcal{L}^{\mathrm{SoCo}} = \mathcal{L} + \widetilde{\mathcal{L}}.$$



Scale-Aware Assignment

- FPN original

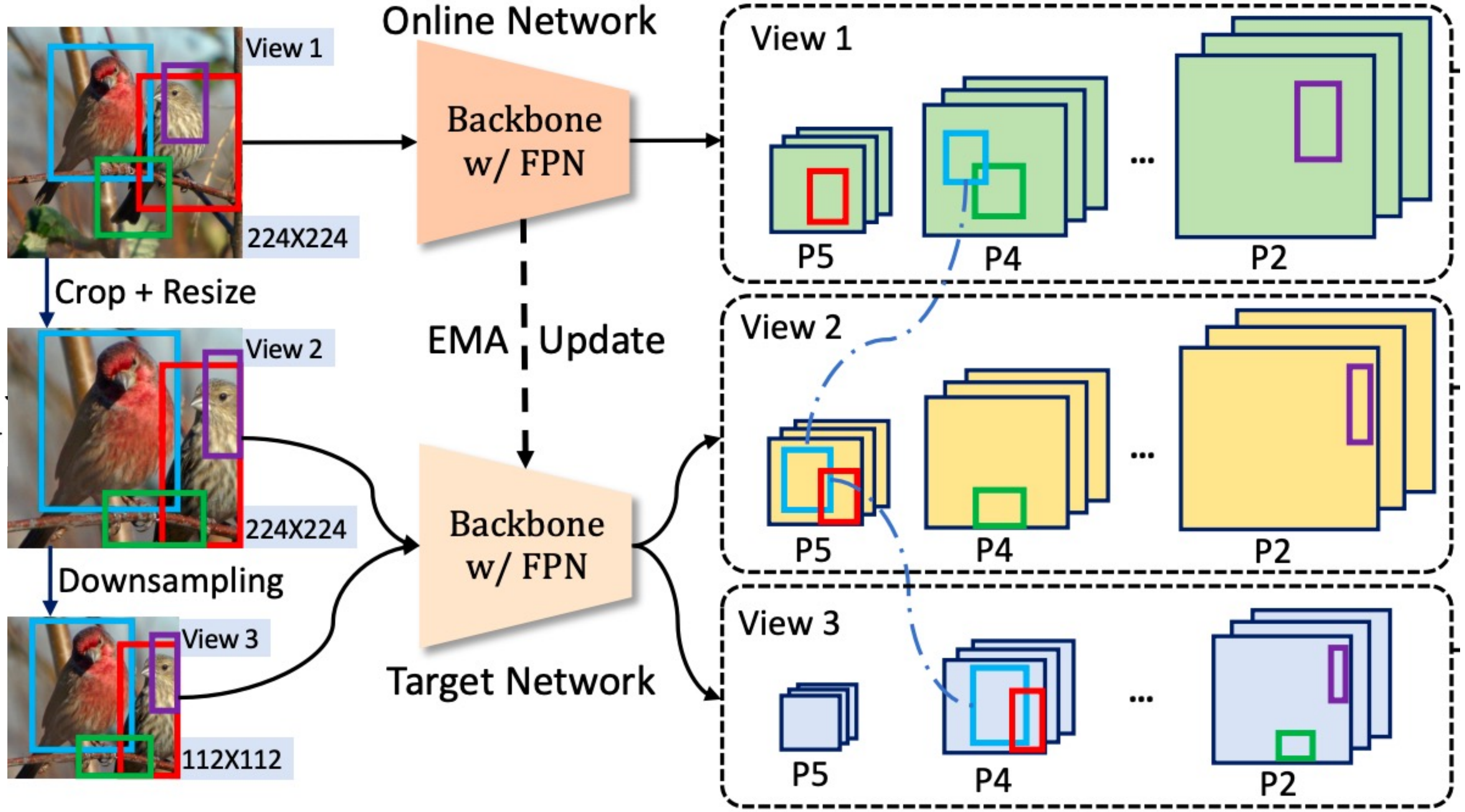
  $\{32^2, 64^2, 128^2, 256^2\}$

  $\{P_2, P_3, P_4, P_5\}$

- In paper

$\{0 - 48^2, 49^2 - 96^2, 97^2 - 192^2, 193^2 - 224^2\}$

$\{P_2, P_3, P_4, P_5\}$



Result

Table 1: Comparison with state-of-the-art methods on **COCO** by using Mask R-CNN with **R50-FPN**.

Methods	Epoch	1× Schedule						2× Schedule					
		AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
Scratch	-	31.0	49.5	33.2	28.5	46.8	30.4	38.4	57.5	42.0	34.7	54.8	37.2
Supervised	90	38.9	59.6	42.7	35.4	56.5	38.1	41.3	61.3	45.0	37.3	58.3	40.3
MoCo [4]	200	38.5	58.9	42.0	35.1	55.9	37.7	40.8	61.6	44.7	36.9	58.4	39.7
MoCo v2 [5]	200	40.4	60.2	44.2	36.4	57.2	38.9	41.7	61.6	45.6	37.6	58.7	40.5
InfoMin [6]	200	40.6	60.6	44.6	36.7	57.7	39.4	42.5	62.7	46.8	38.4	59.7	41.4
BYOL [3]	300	40.4	61.6	44.1	37.2	58.8	39.8	42.3	62.6	46.2	38.3	59.6	41.1
SwAV [7]	400	-	-	-	-	-	-	42.3	62.8	46.3	38.2	60.0	41.0
ReSim-FPN^{T} [41]	200	39.8	60.2	43.5	36.0	57.1	38.6	41.4	61.9	45.4	37.5	59.1	40.3
PixPro [10]	400	41.4	61.6	45.4	-	-	-	-	-	-	-	-	-
InsLoc [12]	400	42.0	62.3	45.8	37.6	59.0	40.5	43.3	63.6	47.3	38.8	60.9	41.7
DenseCL [11]	200	40.3	59.9	44.3	36.4	57.0	39.2	41.2	61.9	45.1	37.3	58.9	40.1
DetCon$_S$ [13]	1000	41.8	-	-	37.4	-	-	42.9	-	-	38.1	-	-
DetCon$_B$ [13]	1000	42.7	-	-	38.2	-	-	43.4	-	-	38.7	-	-
SoCo	100	42.3	62.5	46.5	37.6	59.1	40.5	43.2	63.3	47.3	38.8	60.6	41.9
SoCo	400	43.0	63.3	47.1	38.2	60.2	41.0	44.0	64.0	48.4	39.0	61.3	41.7
SoCo*	400	**43.2**	**63.5**	**47.4**	**38.4**	**60.2**	**41.4**	**44.3**	**64.6**	**48.9**	**39.6**	**61.8**	**42.5**

$\star$ Additional V4, 192 x 192

Ablation Study

Table 4: Ablation study on the effectiveness of aligning pretraining to object detection.

Whole Image	Selective Search	FPN	Head	Scale-aware Assignment	Box Jitter	Multi View	AP^{bb}	AP^{mk}
✓							38.1	34.4
✓	✓						40.6 (+2.5)	36.8 (+2.4)
✓	✓	✓					40.2 (+2.1)	36.2 (+1.8)
✓	✓	✓	✓				41.2 (+3.1)	37.0 (+2.6)
✓	✓	✓	✓	✓			41.6 (+3.5)	37.3 (+2.9)
✓	✓	✓	✓	✓	✓		41.7 (+3.6)	37.5 (+3.1)
✓	✓	✓	✓	✓	✓	✓	**42.3 (+4.2)**	**37.6 (+3.2)**

Ablation Study

Table 5: Ablation studies on hyper-parameters for the proposed SoCo method.

(a) Study on image size of view V_3.

Image Size	AP^{bb}	AP^{mk}
96	42.1	37.7
112	**42.3**	37.6
128	42.1	37.7
160	42.0	37.6
192	42.2	**37.8**

(b) Study on batch size.

Batch Size	AP^{bb}	AP^{mk}
512	41.7	**37.6**
1024	41.9	**37.6**
2048	**42.3**	**37.6**
4096	41.4	37.3

(c) Study on proposal generation and proposal number K.

Selective Search	Random	K	AP^{bb}	AP^{mk}
✓		1	41.6	37.3
✓		**4**	**42.3**	**37.6**
✓		8	41.6	37.4
✓		16	41.2	37.0
	✓	1	41.4	36.9
	✓	4	NaN	NaN
	✓	8	NaN	NaN

(d) Study on momentum coefficient τ.

τ	AP^{bb}	AP^{mk}
0.98	35.0	31.7
0.99	**42.3**	**37.6**
0.993	41.8	**37.6**