CMFPN: Context Modeling Meets Feature Pyramid Network

Faroq AL-Tam, Muhammed AL-Qurishi, Thariq Khalid, Riad Souissi

ELM Company, Riyadh

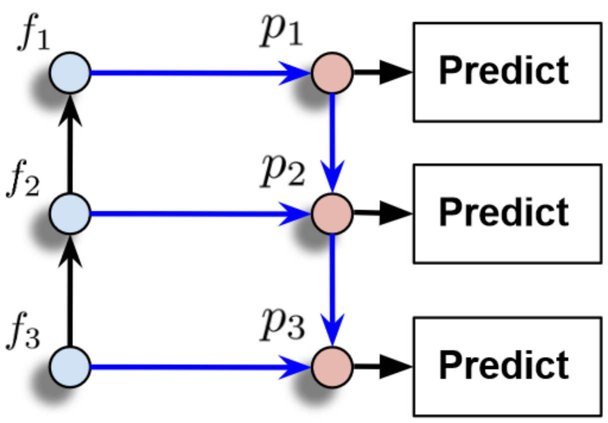
8th of May 2024

ICLR - Vienna

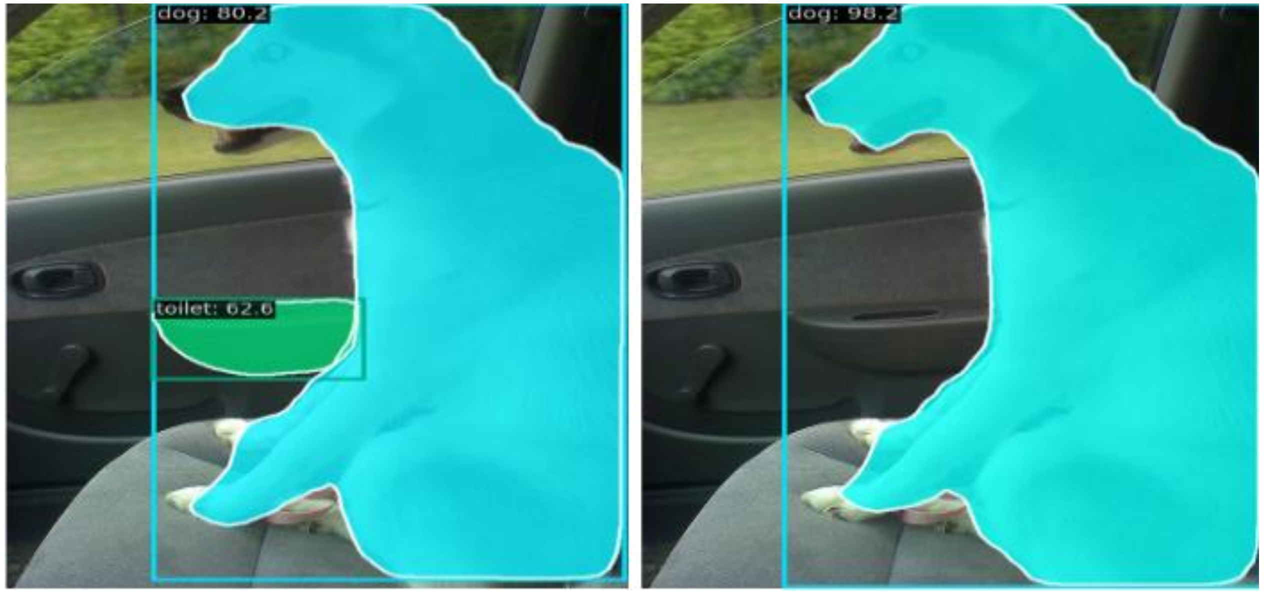
Content

- Motivation
- CMFPN
  - FPN to CMFPN
  - Formulation
- Results
- Conclusion and forward
- Questions

Motivation

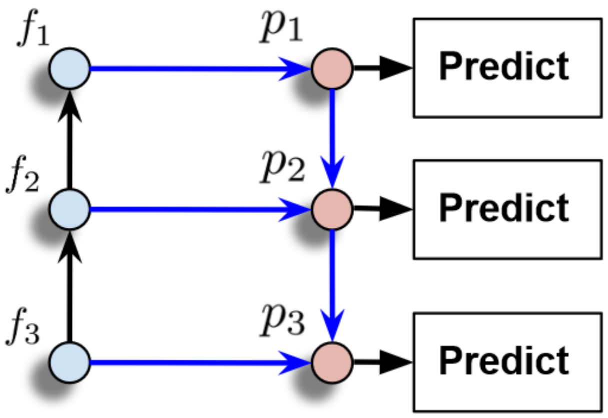


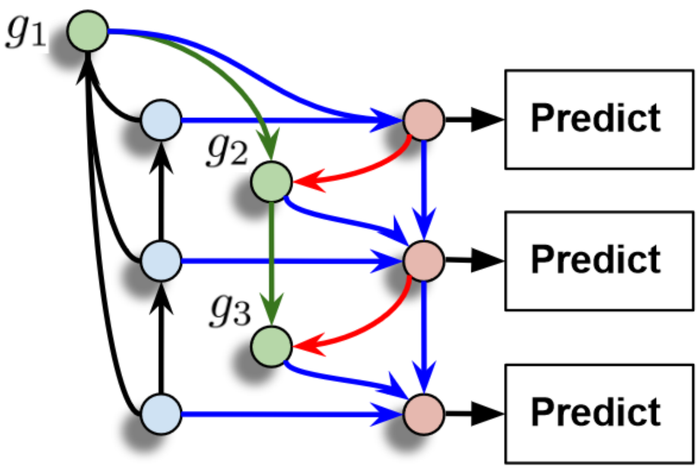
f_1
p_1
Predict
f_2
p_2
Predict
f_3
p_3
Predict



dog: 80.2
toilet: 62.6
dog: 98.2

FPN to CMFPN





$$p_k = \begin{cases} V_k(W_k f_k) & \text{if } k = 1, \\ V_k\left(W_k f_k + p_{k-1}\right) & \text{otherwise,} \end{cases} \tag{1}$$

$$p_k = \begin{cases} V_k\left(W_k f_k + g_k\right) & \text{if } k = 1, \\ V_k\left(W_k f_k + g_k + p_{k-1}\right) & \text{otherwise,} \end{cases} \tag{2}$$

CMFPN latent map:

$$p_k = \begin{cases} V_k\left(W_k f_k + g_k\right) & \text{if } k = 1, \\ V_k\left(W_k f_k + g_k + p_{k-1}\right) & \text{otherwise,} \end{cases} \tag{2}$$

Calibrated backbone feature maps:

$$\tilde{f}_k = \mathrm{Scale}_{2(\bar{k}-k)}(\mathrm{SE}(f_k)), \quad k = 1, \cdots, K, \tag{3}$$

Context:

$$g_k = \begin{cases} V_k^g \mathrm{Concat}_{C_{\mathcal{P}}}(\tilde{\mathcal{F}}) & \text{if } k = 1, \\ V_k^g \left(W_k \tilde{f}_{k-1} + \mathrm{CCM}(g_{k-1}, p_{k-1}) + g_{k-1}\right) & \text{otherwise,} \end{cases} \tag{4}$$

Context meets latent maps:

$$\mathrm{CCM}(g_k, p_k) = V_k^{\mathrm{CM}} \mathrm{CM}\left(\mathrm{Concat}_{C_{\mathcal{P}}}\left(g_k, \mathrm{Scale}_{2(\bar{k}-k)}(p_k)\right)\right), \tag{5}$$

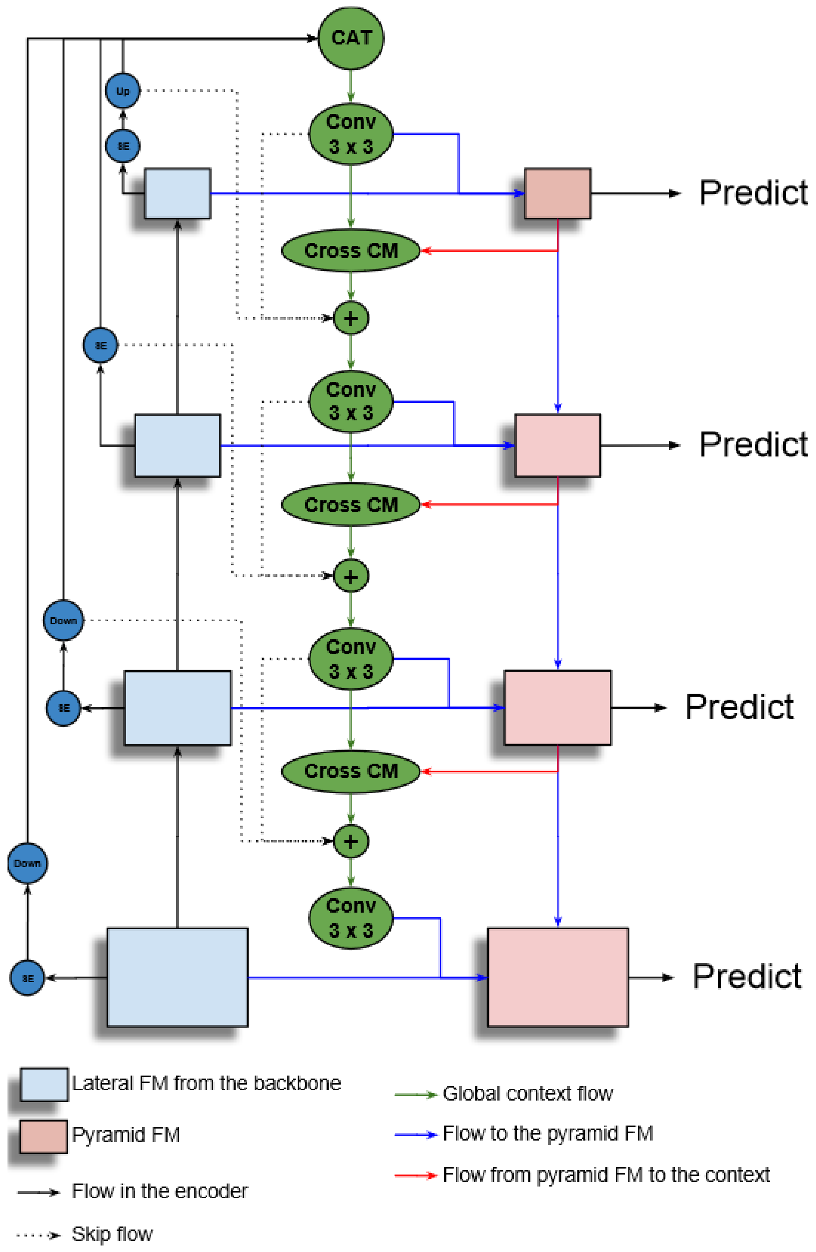
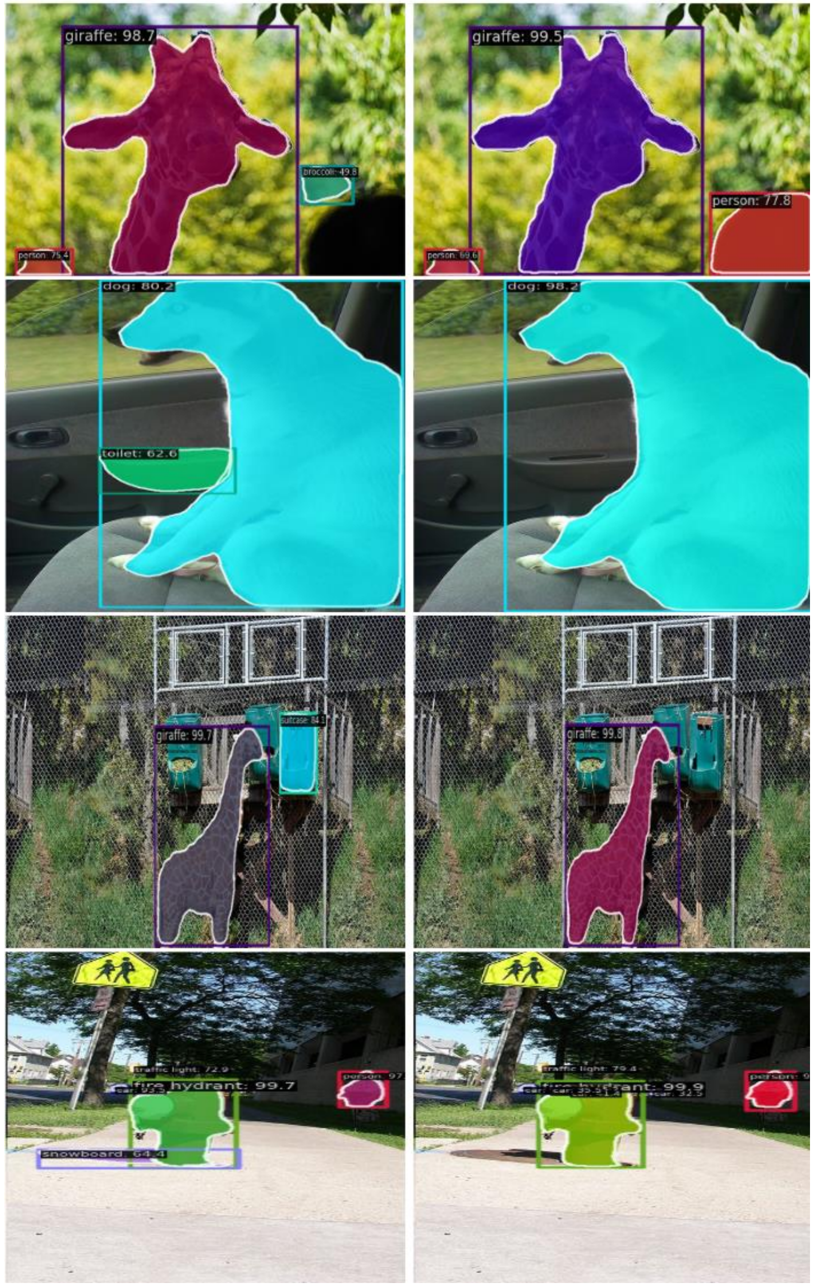


Figure 3: CMFPN


giraffe: 98.7
giraffe: 99.5
person: 77.8
dog: 80.2
dog: 98.2
toilet: 62.6
giraffe: 99.7
giraffe: 99.8
fire hydrant: 99.7
snowboard: 64.4
fire hydrant: 99.9

FPN CMFPN

Results (OD)

Model	**Backbone**	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster R-CNN	R-50 + FPN	36.90	58.40	39.70	21.70	40.50	48.10
Faster R-CNN	R-50 + CFPN Xie et al. (2023)	37.20	-	-	21.70	41.40	48.60
YOLOF	R-50 Chen et al. (2021)	37.70	56.90	40.60	19.10	42.5	53.20
Faster R-CNN	R-50 + CMFPN	$39.00_{(+2.1)}$	60.50	42.30	22.90	42.20	51.60
Mask R-CNN	R-50 + FPN	37.40	58.50	40.10	21.70	40.70	48.60
Mask R-CNN	R-50 + CMFPN	$39.60_{(+2.2)}$	60.90	42.90	23.80	43.00	52.40
Cascade Mask R-CNN	R-50 + FPN	40.70	59.10	44.30	22.50	44.30	54.00
Cascade Mask R-CNN	R-50 + CFPN Xie et al. (2023)	41.50	-	-	24.10	45.70	54.00
Cascade Mask R-CNN	R-50 + CMFPN	$42.90_{(+2.2)}$	62.00	46.40	25.40	46.60	57.10
Mask R-CNN	Swin-T + FPN	42.40	65.10	46.10	25.80	45.60	56.10
Mask R-CNN	Swin-T + CMFPN	$45.10_{(+2.7)}$	67.00	48.90	27.30	48.80	60.40

Results (IS)

Model	**Backbone**	AP^{Seg}	AP^{Seg}_{50}	AP^{Seg}_{75}	AP^{Seg}_{S}	AP^{Seg}_{M}	AP^{Seg}_{L}
Mask R-CNN	R-50 + FPN	33.90	55.10	36.00	16.00	36.50	49.80
Mask R-CNN	R-50 + CMFPN	35.60$_{(+1.7)}$	57.50	37.70	17.50	38.20	51.90
Cascade Mask R-CNN	R-50 + FPN	35.30	56.00	37.80	16.20	38.00	51.80
Cascade Mask R-CNN	R-50 + CMFPN	37.10$_{(+1.8)}$	58.50	39.70	18.40	39.80	54.30
Mask R-CNN	Swin-T + FPN	39.10	62.10	42.10	19.60	41.80	57.50
Mask R-CNN	Swin-T + CMFPN	40.70$_{(+1.6)}$	64.20	43.70	21.00	43.80	60.00

Table 2: The instance segmentation results on the coco *val* 2017.

Conclusions

- FPN fuses multiscale features but it brings suboptimal context to the detection heads.
- CMFPN resolves these issues by modeling the context separately.
- Results show consistent performance on different backbones and object sizes.
- CMFPN will be extended by novel context-aware selective attention.

Thank you

faroq.al.tam@gmail.com