

# Detekcijska metoda za štetje objektov z malo učnimi primeri

Jer Pelhan, Alan Lukežič, Vitjan Zavrtnik, Matej Kristan

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Večna pot 113, 1000 Ljubljana  
E-pošta: jer.pelhan@gmail.com, {alan.lukezic, vitjan.zavrtanik, matej.kristan}@fri.uni-lj.si

## Towards few-shot counting by detection<sup>1</sup>

*While few-shot counting has been dominated by global counters, these do not provide object locations, crucial for some applications. This was addressed by repurposing detector architectures for the task, however detection-based counters still lag behind in global counts. We propose a detection-based counter, which revisits classical detection approaches in the context of modern few-shot counting. Our results indicate that carefully repurposing global counters may lead to methods that challenge the performance of state-of-the-art detection-based counters, thus opening the potential for combining the best properties of the two counting paradigms.*

## 1 Uvod

Štetje z malo učnimi primeri [13] naslavlja problem štetja objektov v sliki, kjer imamo podanih zgolj nekaj (tipično tri) učnih primerov. Problem izhaja iz širšega področja štetja, ki se v glavnem fokusira na štetje v naprej znanih kategorij, kot so ljudje [9, 16] in vozila [2]. Poseben izziv štetja z malo učnimi primeri je sposobnost gradnje robustnega modela v delno-nadzorovanem učnem režimu zgolj na podlagi ene slike.

Trenutno najuspešnejše, tako imenovane globalne metode, temeljijo na gradnji korelacijskih filtrov, s katerimi se korelira vhodna slika (v prostoru globokih značilnk). Število objektov se oceni z vsoto preko dobljene korelacijske mape. Izziv je torej učenje filtrov, ki dobro generalizirajo preko izgledov objektov v sliki. Slednji problem se v literaturi naslavlja s tehnikami učenja metrične podobnosti [15], izboljšane metode korelacije značilnk prototipov [22] ali interakcije značilnk, ki temeljijo na transformerjih [7].

Globalne metode podajo zgolj število objektov, ne pa tudi njihove lokacije, kar je pomembna informacija za številne aplikacije, kot so na primer biološke analize [25]. V zadnjem času so se zato pojavile metode temelječe na detektorjih [12], vendar pa slednje ne dosegajo natančnosti štetja globalnih metod. Skupni problem vseh metod, pa je v konfliktu med priklicem in natančnostjo. Metode namreč močno generalizirajo preko izgledov objektov za povečevanje priklica. Slednje pa rezultira v

nižji natančnosti, saj generalizacija posplošuje tudi preko napačnih pozitivov, t.j., drugih prisotnih objektov.

Spričo omenjenih izzivov predlagamo novo detekcijsko metodo za štetje objektov TCD (angl., two-stage counting by detection). TCD je dvostopejska metoda, ki naslavlja problem konflikta med priklicem in natančnostjo. V prvem koraku TCD maksimizira priklic tako, da generira potencialno redundantno število regij s kandidati objektov, v naslednjem koraku pa poveča natančnost z verifikacijo kandidatov. Rezultati kažejo, da TCD prekaša trenutno najboljšo metodo za štetje z detekcijo.

## 2 Sorodna dela

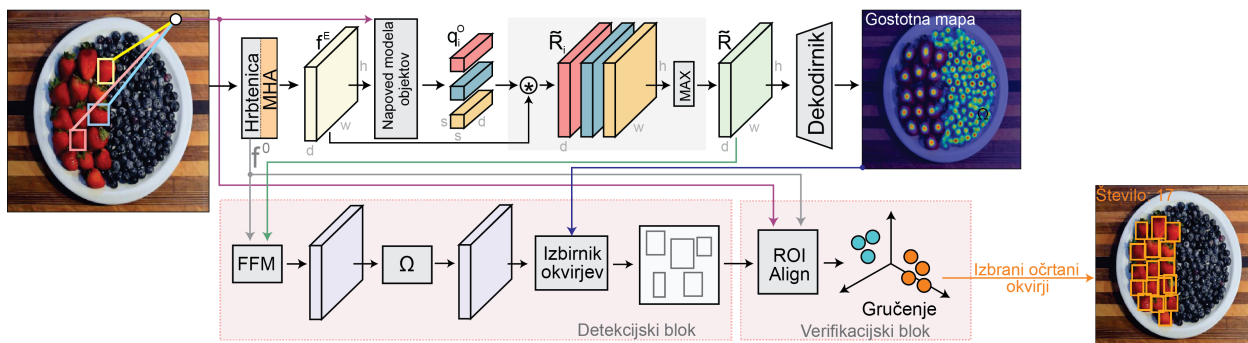
Sodobne metode štetja z malo učnimi primeri v glavnem temeljijo na ekstrakciji značilnk primerkov in primerjavo le-teh s celotno sliko. Metoda Famnet [13] značilke primerkov uporabi kot korelacijska jedra, kar sledi formulaciji siamskih sledilnikov [1]. Avtorji metode BMNet+ [15] korelacijo iz [13] zamenjajo z bilinearno učljivo mero podobnosti. Z mehanizmom pozornosti metoda še dodatno zajame variabilnost med objekti v sliki. Metoda SAFECount [22] naslovi problem slik z velikim številom objektov tako, da s pomočjo podobnostne mape izračuna uteži za združitev značilnk primerkov in slike. CounTR [8] za ekstrakcijo značilnk primerkov uporablja konvolucijsko nevronska mrežo, za ekstrakcijo značilnk slike pa ViT transformerski model [4]. Pri iskanju podobnosti uporabi mehanizem med-pozornosti (angl., cross-attention). Vse opisane metode napovejo zgolj globalno število objektov na sliki, ne pa tudi njihove lokacije.

Pred kratkim predstavljena metoda za štetje in detekcijo z malo učnimi primeri C-DETR [12] temelji na metodi za detekcijo objektov [19]. Glavna slabost takih metod je, da na slikah z velikim številom objektov ne delujejo dobro. Zato metoda C-DETR ne dosega rezultatov primerljivih z metodami za zgolj štetje objektov z malo učnimi primeri.

## 3 Metoda TCD

Ob podani vhodni sliki  $I$  velikosti  $H_0 \times W_0 \times 3$  in naborom  $k$  očitanih okvirjev  $B = \{b_1, b_2, \dots, b_k\}$ , ki zajemajo vizualne značilnosti ciljnega razreda objektov tj.  $k$  primerkov, je cilj metode napovedati število in lokacije objektov, ki pripadajo ciljnemu razredu.

<sup>1</sup>Delno financirano iz ARIS programa P2-0214 in projektov J2-2506 in Z2-4459



Slika 1: Arhitektura dvostopenjske metode za štetje in detekcijo objektov z malo učnimi primeri TCD. Detekcijski blok najprej napove očitane okvirje, med katerimi so potencialno tudi napačni, katere odstrani verifikacijski blok.

### 3.1 Osnovna struktura

Predlagana metoda TCD je dvostopenjska in sledi principu zaznaj in popravi. V prvem delu uporabimo zanesljivo metodo za lokalizacijo objektov z visokim priklicem. Kot hrbtnico metode TCD uporabimo metodo za štetje z malo učnimi primeri LOCA [3], ki oceni približno lokacijo objektov. Na podlagi te ocene detekcijski blok vsakemu zaznanemu objektu, tj. kandidatu, priredi natančen očitani okvir. Sledi verifikacijski blok, ki preveri vsakega izmed kandidatov in ga po potrebi odstrani, ter tako zmanjša število napačnih pozitivnih detekcij in s tem izboljša natančnost metode. Arhitektura predlagane metode TCD je prikazana na Sliki 1.

### 3.2 LOCA

Za hrbtnico predlagane metode TCD smo uporabili nedavno objavljeno metodo za štetje z malo učnimi primeri LOCA [3]. Vhodna slika se najprej preslika v visokodimenzionalni prostor značilk  $\mathbf{f}^E \in \mathbb{R}^{h \times w \times d}$ , kjer so  $h$ ,  $w$  in  $d$  višina, širina in število kanalov značilk. Glede na lokacije označenih  $k$  primerkov LOCA v prostoru značilk zgradi  $k$  korelacijskih filtrov, ki jih korelira z mapo značilk  $\mathbf{f}^E$  in tako dobi mapo podobnosti  $\tilde{\mathbf{R}} \in \mathbb{R}^{h \times w \times d}$ . S pomočjo dekodirnika preslika  $\tilde{\mathbf{R}}$  v dvo-dimenzionalno gostotno mapo  $\mathbf{G} \in \mathbb{R}^{H_0 \times W_0}$ . Vsota vseh elementov v  $\mathbf{G}$  predstavlja končno število objektov izbranega razreda v sliki.

### 3.3 Detekcijski blok

Za napoved očitanih okvirjev uporabimo značilke iz celotne slike  $\mathbf{f}^O \in \mathbb{R}^{h \times w \times c}$ , ki jih LOCA izračuna v zgodnjih fazah cevovoda. Te značilke vsebujejo informacijo o izgledu objektov in jih najprej združimo s podobnostno mapo  $\tilde{\mathbf{R}}$ , ki vsebuje informacijo o približni lokaciji objektov. Zdržitev poteka z uporabo FFM združitevne modula [23], t.j.,  $\tilde{\mathbf{f}} = \text{FFM}(\mathbf{f}^O, \tilde{\mathbf{R}})$ , kjer je  $\tilde{\mathbf{f}} \in \mathbb{R}^{H_0 \times W_0 \times d}$ .

Napoved lokacije objektov se izvede na vsaki poziciji  $(x, y)$  znotraj  $\tilde{\mathbf{f}}$  z uporabo transformacije:  $\mathbf{v} = \Omega(\tilde{\mathbf{f}})$ , kjer je  $\mathbf{v} \in \mathbb{R}^{H_0 \times W_0 \times 4}$ . Uporabimo t.i., *tlrb* zapis lokacije [24], ki je definiran kot:  $\mathbf{v}(x, y) = [t^*, l^*, r^*, b^*]$ . Očitani okvir  $b_i$  je definiran kot  $b_i = (x_0^{(i)}, y_0^{(i)}, x_1^{(i)}, y_1^{(i)})$ , kjer sta  $(x_0^{(i)}, y_0^{(i)})$  in  $(x_1^{(i)}, y_1^{(i)})$  koordinata zgornjega levega in spodnjega desnega oglišča.

Koordinate izračunamo s pomočjo naslednje enačbe:

$$\begin{aligned} x_0^{(i)} &= x - l^*, & x_1^{(i)} &= x + r^*, \\ y_0^{(i)} &= y - t^*, & y_1^{(i)} &= y + b^*. \end{aligned} \quad (1)$$

Transformacija  $\Omega$  je definirana z naslednjim zaporedjem operacij:  $3 \times 3$  conv ( $d \rightarrow d$ ), GN, ReLU,  $3 \times 3$  conv ( $d \rightarrow 4$ ). Operacija  $3 \times 3$  conv ( $n_{\text{IN}} \rightarrow n_{\text{OUT}}$ ) predstavlja 2-dimenzionalno konvolucijo z jedri velikosti  $3 \times 3$ , ki vhod z  $n_{\text{IN}}$  kanali preslika v izhod z  $n_{\text{OUT}}$  kanali. GN predstavlja operacijo skupinske normalizacije [20] (angl., group norm), ReLU pa usmerjeno linearno enoto.

Prejšnji korak nam na vsaki lokaciji  $(x, y)$  v sliki napove očitani okvir ne glede na to, kaj se na tej lokaciji nahaja. Da bi odstranili napovedi, ki pripadajo ozadju vpeljemo izbirnik okvirjev, ki izbere samo tiste očitane okvirje, ki pripadajo objektom, glede na gostotno mapo  $\mathbf{G}$ , ki jo napove LOCA. Izbirnik okvirjev najprej postavi na 0 vse vrednosti v  $\mathbf{G}$ , ki niso dovolj visoke:

$$\mathbf{G}^*(x, y) = \begin{cases} 0, & \mathbf{G}(x, y) < 0.25 \cdot \text{MAX}(\mathbf{G}) \\ \mathbf{G}(x, y), & \text{drugače,} \end{cases}$$

pri čemer  $\text{MAX}(\cdot)$  predstavlja element z najvišjo vrednostjo. Sledi korak zadužitve vrednosti, ki niso lokalni maksimumi:

$$\mathbf{G}^{**}(x, y) = \begin{cases} \mathbf{G}^*(x, y), & \mathbf{G}^*(x, y) \gg \kappa(\mathbf{G}^*(x, y)) \\ 0, & \text{drugače,} \end{cases}$$

kjer je  $\kappa(\cdot(x, y))$  operacija, ki vrne  $r \times r$  elementov okoli lokacije  $(x, y)$ , simbol  $\gg$  pa predstavlja neenakost, ki je resnična, ko je na levi strani element, ki je večji od vseh elementov na desni strani neenakosti. Rezultat detekcijskega bloka je množica očitanih okvirjev, kandidatov objektov,  $B^{\text{DET}} = \{b_i\}_{i=1:m}$  na vseh lokacijah  $(x, y)$ , kjer velja  $\mathbf{G}^{**}(x, y) > 0$ .

### 3.4 Verifikacijski blok

Detekcijski blok zaradi prekomerne generalizacije med kandidati za objekte pogosto predlaga tudi napačne pozitive, tj. objekte, ki ne pripadajo izbrani kategoriji za štetje. V drugem koraku, zato predlagamo verifikacijski blok, ki temelji na gručenju in iz množice kandidatov  $B^{\text{DET}}$  odstrani detekcije, ki z veliko verjetnostjo ne pripadajo izbrani kategoriji.

Verifikacijski blok najprej izlušči značilke iz celotne slike  $\mathbf{f}^0$  na  $m$  lokacijah  $b_i \in B^{\text{DET}}$  s pomočjo operacije RoiAlign [6]. Na enak način izlušči tudi značilke  $k$  primerkov. Rezultat je množica  $m + k$  značilk, kjer je vsaka velikosti  $3 \times 3 \times s$ . Nato med vsemi pari značilk izračunamo kosinusno podobnost, ki jo predstavimo z matriko  $\mathcal{S} \in \mathbb{R}^{(m+k) \times (m+k)}$ . Z uporabo hevrstike lastnih vrzeli [17] (angl., Eigen gap) ocenimo število gruč v matriki  $\mathcal{S}$ , ter izvedemo spektralno gručenje podatkov [11] (angl., Spectral clustering), ki vsakemu izmed  $m + k$  podatkov priredi oznako kateri gruči pripada.

Med predlaganimi gručami, za pozitivne oziroma ciljne označimo vse, ki vsebujejo vsaj en primer, ostale označimo kot negativne gruče. V množici  $B^{\text{DET}}$  ohranimo očitane okvirje, ki so pripadniki pozitivnih gruč, ostale pa zavržemo. Tako dobimo končne detekcije, njihovo število pa predstavlja oceno števila objektov označene kategorije.

### 3.5 Postopek učenja

Parametri hrbtenice štetja LOCA, so pred-naučeni in se v nobeni od učnih faz ne spreminjajo. Detekcijski in verifikacijski blok pa učimo v dveh ločenih fazah.

V prvi fazi učimo natančno napovedovanje očitanih okvirjev v detekcijskem bloku (Podpoglavje 3.3) z minimizacijo naslednje kriterijske funkcije:

$$\mathcal{L}_1 = \sum_{i=1}^k 1 - \text{GIoU}(\mathbf{v}(x_i^c, y_i^c), b_i^{\text{GT}}), \quad (2)$$

kjer je  $\mathbf{v}(x_i^c, y_i^c)$  napovedan očitani okvir na lokaciji  $(x_i^c, y_i^c)$ , ki predstavlja središče  $i$ -tega označenega očitane okvirja  $b_i^{\text{GT}}$  in GIoU funkcija generaliziranega preseka nad unijo [14] (angl., Generalized IoU). Poudarimo, da učenje poteka zgolj na  $k$  označenih očitanih okvirjih, saj so to edini nadzorovani učni primeri, ki so na voljo.

V drugi fazi učimo samo verifikacijski blok (Podpoglavje 3.4) za maksimizacijo sposobnosti razlikovanja med izgledi objektov različnih kategorij. Zato posamezen učni primer kreiramo tako, da staknemo dve sliki, pri čemer preprečimo združevanje slik, ki vsebujejo objekte iste kategorije. Nato iz vsake od združenih slik izluščimo značilke, ki pripadajo označenim objektom, kot je opisano v Podpoglavju 3.4. Dobimo dve množici značilk, kjer  $\{\mathbf{z}_j^1\}_{j=1:k}$  pripada prvi sliki, množica  $\{\mathbf{z}_j^2\}_{j=1:k}$  pa drugi sliki. Učenje je definirano kot minimizacija kontrastne kriterijske funkcije [18]:

$$\mathcal{L}_2 = \sum_{i_1=1}^2 \sum_{j_1=1}^k \sum_{i_2=1}^2 \sum_{j_2=1}^k \ell(z_{j_1}^{i_1}, z_{j_2}^{i_2}), \quad (3)$$

kjer je funkcija  $\ell(\cdot, \cdot)$  definirana kot:

$$\ell(z_{j_1}^{i_1}, z_{j_2}^{i_2}) = \begin{cases} 1 - d(z_{j_1}^{i_1}, z_{j_2}^{i_2}), & i_1 = i_2 \\ \max(0, d(z_{j_1}^{i_1}, z_{j_2}^{i_2}) - \lambda), & \text{drugače.} \end{cases}$$

Funkcija  $d(z_{j_1}^{i_1}, z_{j_2}^{i_2})$  predstavlja kosinusno podobnost med parom značilk,  $\lambda$  pa je t.i., mejni parameter kontrastnega kriterijske funkcije (angl., margin).

## 4 Eksperimenti

### 4.1 Evalvacijski protokol

Predlagano metodo TCD evalviramo na podatkovni zbirki za štetje in detekcijo objektov z malo učnimi primeri, FSCD-147 [12], ki predstavlja razširitev osnovne zbirke [13]. Zbirka zajema 147 različnih kategorij, slike so razdeljene na sklope za učenje, preverjanje in testiranje, sestavljene iz 3659, 1286 in 1190 slik, število objektov na posamezni sliki pa se giblje med 7 in 3731. Na vsaki sliki so z očitanim okvirjem označeni trije objekti ( $k = 3$ ), ki določajo kategorijo objektov, ki jih želimo prešteti. Vsi objekti ciljne kategorije so označeni z lokacijo središča objekta, v validacijski in testni množici pa tudi z očitanim okvirjem, kar omogoča evalvacijo detekcijskih sposobnosti.

Za ocenjevanje uspešnosti metode za štetje uporabimo standardni meri [10], ki merita odstopanje napovedanega od pravega števila objektov v sliki. To sta povprečna absolutna napaka (angl., Mean Absolute Error – MAE) in korenjena povprečna kvadratna napaka (angl., Root Mean Squared Error - RMSE). Za oceno uspešnosti lokalizacije objektov uporabimo klasični detekcijski meri [12]: povprečna natančnost (AP) in povprečna natančnost pri IoU pragovni vrednosti 0.5 (AP50).

### 4.2 Evalvacija na FSCD-147

Sposobnost štetja objektov metode TCD je ocenjena na podatkovni zbirki FSCD-147 [13], rezultati so podani v Tabeli 1. Naša metoda TCD prekaša vse obstoječe metode za štetje in detekcijo z malo učnimi primeri, vključno z najsodobnejšo metodo C-DETR [12]. V primerjavi z metodo C-DETR namreč dosega za kar 49% in 48% manjšo napako na validacijski množici, ter 32% in 33% na testni množici, po merah MAE in RMSE.

Sposobnost detekcije objektov metode TCD, t.j., natančnost napovedi lokacije objektov ciljne kategorije, je predstavljena v Tabeli 2. Predlagano metodo primerjamo s prilagojeno osnovno metodo za štetje objektov z malo učnimi primeri Famnet [13], z metodama za detekcijo objektov z malo učnimi primeri Attention-RPN [5] in FSDeView [21] in s trenutno najboljšo predlagano metodo za štetje in detekcijo z malo učnimi primeri C-DETR [12]. TCD dosega za 19% in 37% boljšo lokalizacijo na validacijski podmnožici v primerjavi s C-DETR izmerjeno z AP in AP50. TCD prav tako izboljša rezultat tudi na testni podmnožici in sicer za 4% in 20% izmerjeno z AP in AP50. Rezultati torej kažejo na izjemno velik potencial predlagane metode TCD za štetje in detekcijo objektov z malo učnimi primeri.

## 5 Zaključek

Predstavili smo TCD, dvostopenjsko metodo za štetje in detekcijo objektov poljubnih kategorij z malo učnimi primeri. Metoda naslavlja problem konflikta med priklicem in natančnostjo tako, da v prvem koraku maksimizira priklic z generiranjem potencialno redundantnega števila kandidatov, v naslednjem koraku pa poveča natančnost z verifikacijo teh kandidatov. Predlagano metodo smo

Tabela 1: Rezultati štetja na validacijski in testni množici FSC-147 za štetje in detekcijo objektov z malo učnimi primeri. Razširitve primerjalnih metod označenih z \* so navedene v [12].

Metoda	VAL		TEST	
	MAE	RMSE	MAE	RMSE
FamNet [13]+RR*	23.75	69.07	22.08	99.54
FSDetView [21]+PB*	/	/	37.54	147.07
FSDetView [21]+RR*	/	/	37.83	146.56
Att-RPN [5]+RR*	/	/	32.70	141.07
Att-RPN [5]+PB*	/	/	32.42	141.55
C-DETR [12]	20.38	82.45	16.79	123.56
TCD (naša)	10.46	43.27	11.40	82.52

Tabela 2: Rezultati detekcij na validacijski in testni množici FSCD-147 za štetje in detekcijo objektov z malo učnimi primeri. Razširitve primerjalnih metod označenih z \* so navedene v [12].

Metoda	VAL		TEST	
	AP	AP50	AP	AP50
FamNet [13]+RR*	/	/	9.44	29.73
FamNet [13]+MLP*	/	/	1.21	6.12
Att-RPN [5]+RR*	/	/	18.53	35.87
Att-RPN [5]+PB*	/	/	13.41	32.99
FSDetView [21]+PB*	/	/	13.41	32.99
FSDetView [21]+RR*	/	/	17.21	33.70
C-DETR [12]	17.27	41.90	22.66	50.57
TCD (naša)	20.57	57.47	23.58	61.14

evalvirali na FSCD-147 [12]. Rezultati kažejo izjemen potencial metode TCD, saj na testni množici dosega 32% in 33% manjšo napako štetja, po merah MAE in RMSE v primerjavi z najboljši metodo C-DETR. Metoda TCD prav tako v primerjavi s C-DETR za 4% in 20% bolj natančno napove lokacije objektov izmerjeno z AP in AP50.

V nadaljnjih raziskavah planiramo dodatne analize v režimu štetja z enim ali brez primerkov, razširitve na interaktivno štetje in možnost izboljšave štetja z zamenjavo hrbtenice, ki jo trenutno predstavlja LOCA [3], z višje-resolucijsko.

## Literatura

[1] Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: ECCV. pp. 850–865. Springer (2016)

[2] Dai, Z., Song, H., Wang, X., Fang, Y., Yun, X., Zhang, Z., Li, H.: Video-based vehicle counting framework. IEEE Access 7, 64460–64470 (2019)

[3] Djukic, N., Lukezic, A., Zavrtanik, V., Kristan, M.: A low-shot object counting network with iterative prototype adaptation. In: ICCV (2023)

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR. OpenReview.net (2021)

[5] Fan, Q., Zhuo, W., Tang, C.K., Tai, Y.W.: Few-shot object detection with attention-rpn and multi-relation detector. In: CVPR. pp. 4013–4022 (2020)

[6] He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)

[7] Liu, C., Zhong, Y., Zisserman, A., Xie, W.: Countr: Transformer-based generalised visual counting. In: BMVC. BMVA Press (2022)

[8] Liu, C., Zhong, Y., Zisserman, A., Xie, W.: Countr: Transformer-based generalised visual counting. arXiv preprint arXiv:2208.13721 (2022)

[9] Liu, W., Salzmann, M., Fua, P.: Context-aware crowd counting. In: CVPR (June 2019)

[10] Lu, E., Xie, W., Zisserman, A.: Class-agnostic counting. In: Asian conference on computer vision. pp. 669–684. Springer (2018)

[11] Ng, A., Jordan, M., Weiss, Y.: On spectral clustering: Analysis and an algorithm. NeurIPS 14 (2001)

[12] Nguyen, T., Pham, C., Nguyen, K., Hoai, M.: Few-shot object counting and detection. In: ECCV. pp. 348–365. Springer (2022)

[13] Ranjan, V., Sharma, U., Nguyen, T., Hoai, M.: Learning to count everything. In: CVPR. pp. 3394–3403 (2021)

[14] Rezatofghi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union. In: CVPR (2019)

[15] Shi, M., Lu, H., Feng, C., Liu, C., Cao, Z.: Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. In: CVPR. pp. 9529–9538 (June 2022)

[16] Shu, W., Wan, J., Tan, K.C., Kwong, S., Chan, A.B.: Crowd counting in the frequency domain. In: CVPR. pp. 19618–19627 (2022)

[17] Von Luxburg, U.: A tutorial on spectral clustering. Statistics and computing 17, 395–416 (2007)

[18] Wang, F., Liu, H.: Understanding the behaviour of contrastive loss. In: CVPR. pp. 2495–2504 (2021)

[19] Wang, Y., Zhang, X., Yang, T., Sun, J.: Anchor detr: Query design for transformer-based detector. In: AAAI. vol. 36, pp. 2567–2575 (2022)

[20] Wu, Y., He, K.: Group normalization. arXiv:1803.08494 (2018)

[21] Xiao, Y., Lepetit, V., Marlet, R.: Few-shot object detection and viewpoint estimation for objects in the wild. IEEE TPAMI 45(3), 3090–3106 (2022)

[22] You, Z., Yang, K., Luo, W., Lu, X., Cui, L., Le, X.: Few-shot object counting with similarity-aware feature enhancement. In: WACV. pp. 6315–6324 (2023)

[23] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: ECCV. pp. 325–341 (2018)

[24] Yu, J., Jiang, Y., Wang, Z., Cao, Z., Huang, T.: Unitbox: An advanced object detection network. In: ACM MM. pp. 516–520 (2016)

[25] Zavrtanik, V., Vodopivec, M., Kristan, M.: A segmentation-based approach for polyp counting in the wild. Engineering Applications of Artificial Intelligence 88, 103399 (2020)