DCA-MoE: Spatially Adaptive Cross-Layer Fusion and Density-Routed Experts for Crowd Counting

Anonymous authors
Paper under double-blind review

Abstract

Crowd counting must estimate local density despite large changes in perspective, head scale, occlusion, and background texture. Many multi-level decoders combine backbone features with spatially invariant weights and then apply the same receptive-field pattern everywhere. We introduce DCA-MoE, a decoder that makes both choices position dependent while retaining a frozen DINOv3 encoder. Spatially Adaptive Layer Fusion (SALF) predicts local weights over four aligned backbone features. Density-Routed Multi-Receptive-Field Experts (DR-MoE) then blends local, mid-range, and large-context residual experts through a dense spatial router. An enhanced blockwise-classification head reconstructs density, and DMCount supervision and a routing-balance term train the decoder. On the NWPU-Crowd validation split, the paired ViT-L/16 configuration obtains 31.7 MAE and 72.2 RMSE. In a matched ViT-B/16 study, the complete model records 32.2/75.9, compared with separately selected minima of 36.2/84.7 for fixed fusion and a shared decoder. Most component controls currently use one seed and retain independently selected minima. The results support spatial adaptation in the decoder while motivating fully paired multi-seed evaluation.

1 Introduction

Crowd counting estimates the number and spatial distribution of people in an image. It supports public-space monitoring and transportation planning, but it is difficult because a single scene may contain isolated large heads, strongly occluded small heads, severe perspective variation, and background patterns that resemble crowds. ShanghaiTech, UCF-QNRF, and NWPU-Crowd expose this heterogeneity across scene scale and density (Zhang et al., 2016; Idrees et al., 2018; Wang et al., 2021).

The field has progressed from multi-column scale-specific regressors (Zhang et al., 2016) to dilated convolution (Li et al., 2018), point-supervised distribution matching (Ma et al., 2019; Wang et al., 2020), and direct localization (Song et al., 2021; Liang et al., 2022). These advances improve the output objective and supervision. Another question remains less directly addressed: which representation depth and spatial context should be used at each image location? Shallow features can retain localization detail, deep features provide semantics and broad context, and their usefulness changes between sparse and congested regions.

Feature pyramids and crowd-counting decoders combine multiple levels (Lin et al., 2017; Ma et al., 2022); dynamic counter mixtures and dynamic convolution adapt processing to input content (Wang et al., 2023; Chen et al., 2020). Nevertheless, common fusion operators use addition, concatenation, or globally shared attention, producing one feature-depth preference for an entire image. A following shared convolution also imposes one receptive field on every location. These two fixed decisions can be restrictive when crowd appearance changes sharply within a scene.

Large self-supervised encoders make local feature selection especially useful. Vision Transformers expose layer-wise token features (Dosovitskiy et al., 2021), and DINO, DINOv2, and DINOv3 learn transferable dense representations without task labels (Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2025). Freezing such an encoder lowers the number of trainable parameters and preserves its representation, while transferring more responsibility to the task decoder.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We propose DCA-MoE, illustrated in Figure 1. SALF aligns four intermediate features and predicts a normalized feature-depth distribution at every output position. DR-MoE predicts a second spatial distribution over three residual experts with complementary receptive fields. The router receives fused visual features; its density awareness is induced by downstream density and counting supervision rather than an explicit density-map input. An EBC-style head (Ma et al., 2025) maps the routed features to local count-bin probabilities, whose expectation reconstructs density.

Our contributions are threefold. First, SALF replaces spatially invariant cross-layer summation with local, normalized feature-depth selection. Second, DR-MoE separates feature-depth selection from receptive-field selection and uses a balancing objective to discourage persistent expert collapse. Third, we integrate both modules with frozen DINOv3 ConvNeXt and ViT encoders and report paired NWPU results and a matched single-seed ViT-B/16 component study. We explicitly separate controlled ablation evidence from comparisons that change pretraining, backbone size, or checkpoint rules.

2 Related Work

Crowd counting. MCNN uses parallel columns to handle scale variation (Zhang et al., 2016), whereas CSRNet replaces explicit columns with dilated convolution (Li et al., 2018). Bayesian Loss treats point annotations as probabilistic contributions (Ma et al., 2019), and DMCount matches predicted density to points through optimal transport while preserving global count consistency (Wang et al., 2020). P2PNet and CLTR directly predict person locations (Song et al., 2021; Liang et al., 2022), while STEERER selectively transfers representations across scales (Han et al., 2023). Our contribution is orthogonal to these supervision paradigms: we use DMCount-based supervision to study adaptive fusion and receptive-field selection.

Classification and foundation-model counting. Blockwise classification quantizes local counts (Liu et al., 2020), while spatial divide-and-conquer recursively partitions high-count regions (Xiong et al., 2019). CLIP transfers web-scale image-language alignment (Radford et al., 2021); CrowdCLIP and CLIP-Count use it for weakly supervised or zero-shot counting (Liang et al., 2023; Jiang et al., 2023). CLIP-EBC combines prompt-tuned CLIP features with enhanced blockwise classification and local density reconstruction (Ma et al., 2025). We retain the structured EBC target but use frozen DINOv3 features and focus on spatially varying layer and context selection.

Adaptive fusion and mixture of experts. Feature pyramids established multi-level fusion for dense prediction (Lin et al., 2017); FusionCount and multifaceted attention similarly aggregate crowd features across scales (Ma et al., 2022; Lin et al., 2022). Dynamic Mixture of Counter Network selects among counter branches (Wang et al., 2023), and dynamic convolution predicts input-conditioned kernel attention (Chen et al., 2020). Mixture-of-experts models route inputs to specialized subnetworks (Shazeer et al., 2017; Fedus et al., 2022), including sparse expert routing for vision Transformers (Riquelme et al., 2021). Our router instead forms a dense soft mixture on a shared spatial grid. All lightweight experts are evaluated; the aim is spatial specialization rather than conditional-computation savings.

3 Method

3.1 Frozen feature extraction and alignment

For an image I, a frozen DINOv3 encoder produces four intermediate representations,

$$\{X_i\}_{i=1}^{4} = f_{\text{DINOv3}}(I). \tag{1}$$

For ConvNeXt we use the four stages at strides 4, 8, 16, and 32. For a ViT we select four Transformer blocks; the default is the final four blocks. A learned 1×1 lateral projection and bilinear interpolation align every level to the output stride of 8:

$$F_i = \text{Resize}_{H/8,W/8}(W_i * X_i + b_i), \quad F_i \in \mathbb{R}^{C \times H/8 \times W/8}. \tag{2}$$

The backbone remains in evaluation mode with gradients disabled. Only the lateral projections and subsequent decoder are optimized.

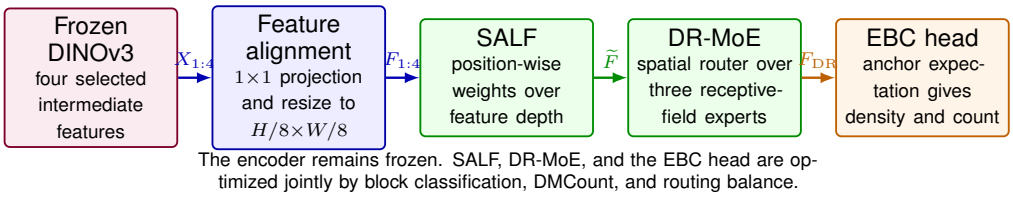

Figure 1: Overview of DCA-MoE. Four frozen DINOv3 features are aligned to a common grid. SALF chooses feature depth per position, DR-MoE chooses receptive-field mixtures per position, and an EBC-style head reconstructs density and image count. The redrawn diagram uses native LaTeX text so labels remain sharp at the ICLR text width.

3.2 Spatially Adaptive Layer Fusion

SALF concatenates the aligned features and predicts four logits per position:

$$A = g_{\mathrm{SALF}}([F_1; F_2; F_3; F_4]), \tag{3}$$

$$\alpha_i(x) = \frac{\exp A_i(x)}{\sum_{j=1}^{4} \exp A_j(x)}. \tag{4}$$

The fused feature is

$$F_{\mathrm{SALF}}(x) = 4 \sum_{i=1}^{4} \alpha_i(x) F_i(x). \tag{5}$$

The factor four preserves the scale of fixed four-feature summation when $\alpha_i = 1/4$. The last gate convolution is initialized to zero, so optimization starts from this uniform state. A 3×3 convolution followed by GroupNorm and GELU produces the smoothed representation $\widetilde{F}$.

3.3 Density-Routed Multi-Receptive-Field Experts

DR-MoE contains three residual experts. E_1 uses a standard 3×3 convolution for local detail; E_2 uses a depthwise 3×3 convolution with dilation two for mid-range context; and E_3 uses a depthwise 7×7 convolution for broader context. Each path includes a 1×1 projection, GroupNorm, GELU, and a residual connection. A router predicts a three-way distribution from $\widetilde{F}$:

$$\pi_e(x) = \mathrm{Softmax}_e(g_{\mathrm{route}}(\widetilde{F})(x)), \tag{6}$$

$$F_{\mathrm{DR}}(x) = \sum_{e=1}^{3} \pi_e(x) E_e(\widetilde{F})(x). \tag{7}$$

Routing is dense and differentiable. The router is not given a predicted density map; it learns density-sensitive decisions through the downstream objective. To reduce collapse, let p_e be the mean routing probability and f_e the fraction of locations whose maximum-probability expert is e. We use

$$\mathcal{L}_{\mathrm{route}} = 3 \sum_{e=1}^{3} p_e f_e. \tag{8}$$

3.4 Density reconstruction and objective

At output reduction eight, the NWPU target uses the intervals $\{[0,0], [1,1], [2,2], [3,3], [4,+\infty)\}$ and anchors $\mathbf{a} = [0, 1, 2, 3, 4.21931]$. Let $z_{b,h,w,k}$ be a class logit and $q_{b,h,w,k}$ its softmax probability. Density and image count are reconstructed as

$$\widehat{D}_{b,h,w} = \sum_{k=1}^{5} q_{b,h,w,k} a_k, \qquad \widehat{y}_b = \sum_{h,w} \widehat{D}_{b,h,w}. \tag{9}$$

The classification term sums pixel-wise cross entropy over the output grid. The density component follows DMCount:

$$\mathcal{L}_{\mathrm{DM}} = \mathcal{L}_{\mathrm{count}} + \lambda_{\mathrm{OT}} \mathcal{L}_{\mathrm{OT}} + \lambda_{\mathrm{TV}} \mathcal{L}_{\mathrm{TV}}. \tag{10}$$

The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{count}}\mathcal{L}_{\text{DM}} + \lambda_{\text{route}}\mathcal{L}_{\text{route}}, \tag{11}$$

with $\lambda_{\text{count}} = 1$, $\lambda_{\text{OT}} = 0.1$, $\lambda_{\text{TV}} = 0.01$, and $\lambda_{\text{route}} = 0.1$ for the full model.

4 Experiments

4.1 Setup

Datasets and metrics. NWPU-Crowd contains 3,109 training, 500 validation, and 1,500 test images (Wang et al., 2021). Its public validation labels support the controlled study; the hidden-label test set is not used for the reported model-selection results. We focus formal evaluation on the labeled NWPU validation split. For N images we use $\text{MAE} = N^{-1}\sum_i |\hat{y}_i - y_i|$ and $\text{RMSE} = \sqrt{N^{-1}\sum_i (\hat{y}_i - y_i)^2}$.

Training. Images preserve aspect ratio, point coordinates are scaled with the image, and sides are adjusted to multiples of 32. Training uses a 448×448 random resized crop with scale range $[0.5, 4.0]$, horizontal flip, mild color jitter, Gaussian blur, and pepper-and-salt noise. We optimize the decoder with AdamW, learning rate 10^{-4}, weight decay 10^{-4}, five warm-up epochs, and cosine decay to one percent of the initial rate. Runs use 2,000 epochs, BF16 mixed precision when available, global batch size 64 on eight GPUs, and seed 42. The ViT-B/16 component study fixes blocks $[8, 9, 10, 11]$ and decoder width 384. Large images are processed with sliding windows; overlapping density predictions are averaged at every output position before summation. MAE selects the checkpoint.

4.2 Comparison with published results

Table 1 places selected paired NWPU validation results beside published baselines. The ViT-B/16 configuration improves over the architecture- matched CLIP-EBC reference by 4.4 MAE and 5.8 RMSE. The paired ViT-L/16 run gives the strongest result in this table. These method comparisons still change pretraining and optimization, so component attribution comes from the matched study below.

4.3 Matched component study

Table 2 fixes the frozen ViT-B/16 backbone, selected layers, decoder width, input pipeline, loss weights, optimizer schedule, validation schedule, and seed. A0 uses fixed summation and one shared decoder. A1 isolates SALF; A2 isolates DR-MoE; A3 combines both without routing balance; and A4 is the complete model.

The complete model is 4.0 MAE and 8.8 RMSE below A0's recorded minima. A1 and A2 alone do not lower both metrics, whereas their combination yields the main change. The balance term further lowers the recorded values from 32.8/77.6 to 32.2/75.9. The selection distinction in Table 2 is important: only A4 currently has both metrics at the same selected checkpoint. Consequently, the table supports the feasibility of combining SALF and DR-MoE but does not yet quantify uncertainty or give a fully paired causal estimate. Repeated seeds and extraction of RMSE at each MAE-selected A0–A3 checkpoint are required before treating the small differences among A0, A1, and A2 as conclusive.

4.4 Backbone, depth, and capacity observations

Scaling is not monotonic. Among paired results, ConvNeXt-Tiny obtains 36.5/94.0, ViT-B/16 obtains 32.2/75.9, and ViT-L/16 obtains 31.7/72.2. Larger ConvNeXt variants in the working records are weaker, but those runs also change decoder width and thus do not isolate backbone capacity. Within fixed-fusion ViT-B/16, selecting the last four blocks $[8, 9, 10, 11]$ improves separately selected minima from 40.5/118.1 for $[1, 8, 10, 11]$ to 36.7/87.2. For ViT-L/16 with width 512, the paired last-four result 31.7/72.2 is better than the cross-depth minima 35.8/115.5. In contrast, ViT-H+/16 records favor cross-depth features under one width. Useful feature depth therefore appears backbone dependent, which motivates SALF, but the current runs do not isolate this factor across every model size.

Table 1: Selected NWPU-Crowd validation results (MAE/RMSE). Published methods use their reported settings; DCA-MoE rows report paired metrics from one checkpoint.

Method	Backbone/setting	MAE/RMSE
P2PNet (Song et al., 2021)	reported	77.4/362.0
DMCount (Wang et al., 2020)	reported	70.5/357.6
CLTR (Liang et al., 2022)	reported	61.9/246.3
mPrompt (Guo et al., 2024)	HRNet-W40-C	50.2/219.0
HMSTUNet (Zhao et al., 2026)	ConvNeXt-S	43.2/119.6
CLIP-EBC (Ma et al., 2025)	ViT-B/16	36.6/81.7
DCA-MoE	ConvNeXt-Tiny	36.5/94.0
DCA-MoE	ViT-B/16	32.2/75.9
DCA-MoE	ViT-L/16	**31.7/72.2**

Table 2: Matched ViT-B/16 component study on NWPU-Crowd. A0–A3 list independently selected MAE and RMSE minima because the source records do not contain paired values at the MAE-selected checkpoint; A4 is paired at epoch 1293. Results use one seed.

Variant	SALF	DR-MoE	Balance	MAE/RMSE	Selection
A0	–	–	–	36.2/84.7	separate
A1	✓	–	–	36.0/87.6	separate
A2	–	✓	✓	36.1/89.1	separate
A3	✓	✓	–	32.8/77.6	separate
A4	✓	✓	✓	**32.2/75.9**	paired

4.5 Qualitative Analysis

Figure 2 combines representative positive and negative cases. The model localizes both a sparse indoor audience and a dense street crowd, with absolute count errors of 0.3 and 12.8. It also suppresses repeated non-human structure: flower rows and gravel produce only 0.7 and 0.2 false counts. The high-count prediction broadly matches the ground-truth support but shows a concentrated high-response region, indicating that count agreement does not imply identical spatial density. These examples illustrate behavior and do not replace split-level metrics or systematic error sampling.

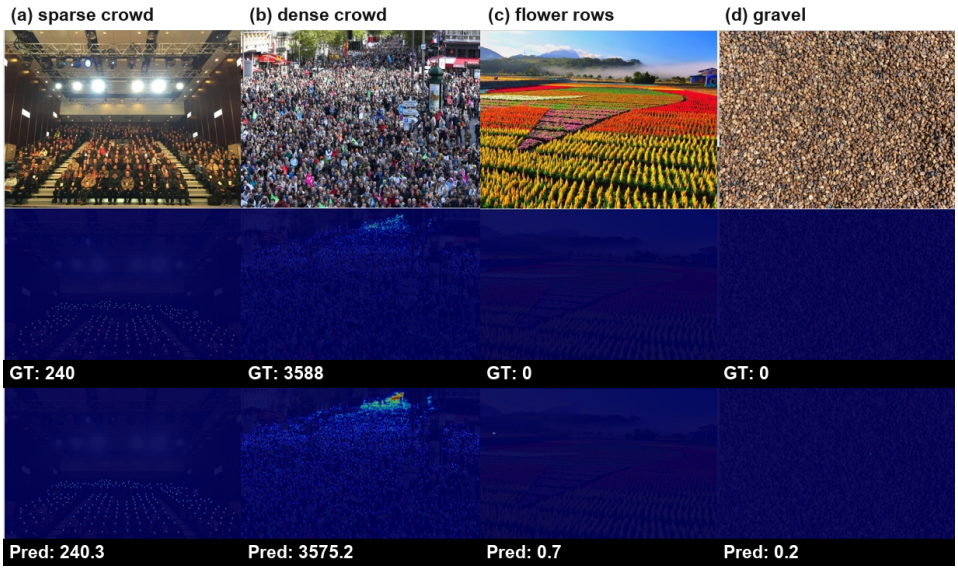

Figure 2: Qualitative cases, with input, ground-truth density, and predicted density from top to bottom. Panels (a,b) show successful sparse and dense counting; panels (c,d) probe false positives from repeated flower rows and gravel. Labels were redrawn at the final column scale for legibility.

5 LIMITATIONS

The present evidence has four limits. First, the central A0–A4 matrix is a single-seed study, and A0–A3 preserve independently selected MAE and RMSE minima from the available records. Second, several backbone comparisons also change decoder width, layer selection, or checkpoint rules. Third, the cross-dataset experiment records in the working draft have incomplete run metadata and are therefore excluded from the formal results. Fourth, "density-routed" means that density supervision shapes a router operating on fused visual features; no explicit density estimate is fed to the router. Paired multi-seed experiments, routing and layer-weight maps, parameter and latency measurements, and matched cross-dataset controls are needed for stronger attribution.

Crowd-counting systems can also be used for surveillance. Dataset bias, privacy, deployment context, and failure under distribution shift should be considered before operational use. Density estimates should support aggregate analysis rather than identification or high-stakes decisions about individuals.

6 CONCLUSION

We presented DCA-MoE, a decoder that adapts feature depth and receptive field at every spatial position while keeping a DINOv3 encoder frozen. SALF learns local cross-layer weights, DR-MoE blends complementary residual experts, and an EBC-style objective reconstructs density and count. Paired NWPU results and a matched component study show promising gains, while incomplete cross-dataset evaluation and single-seed controls prevent a claim of universal superiority. The next priority is a fully paired multi-seed evaluation with routing visualizations and efficiency measurements.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, migration to the ICLR 2027 LaTeX format, and preparation of figure layout and labels. They were not used to generate experimental measurements. The authors reviewed the AI-assisted text, claims, citations, equations, and visual artifacts against the available manuscript and experimental records. The authors take responsibility for the final content, including material produced with AI assistance.

ETHICS STATEMENT

This work uses established crowd-counting benchmarks and does not introduce a new human-subject dataset. Crowd-density estimation may support benign planning and safety applications, but it may also be incorporated into surveillance. Users should consider privacy, dataset bias, distribution shift, local law, and the consequences of errors. The method estimates aggregate density and is not designed for identifying individuals.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 specify the model equations, feature selection, optimization, data processing, metrics, and checkpoint rule. The component study states which factors are held fixed and distinguishes paired from independently selected metrics. The submission package includes the source figures and the script used to rebuild the qualitative montage. Anonymous code, complete run metadata, checkpoint identifiers, and additional paired-seed results should accompany the final submission.

REFERENCES

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11030–11039, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An

image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Mingfei Guo, Li Yuan, Zhi-Qi Yan, Binghui Chen, Yaowei Wang, and Qixiang Ye. Regressor-segmenter mutual prompt learning for crowd counting. *arXiv preprint arXiv:2312.01711*, 2024.

Tao Han, Lei Bai, Lingbo Liu, and Wanli Ouyang. STEERER: Resolving scale variations for counting and localization via selective inheritance learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21848–21859, 2023.

Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *European Conference on Computer Vision*, pp. 532–546, 2018.

Ruixiang Jiang, Lingbo Liu, and Changwen Chen. CLIP-Count: Towards text-guided zero-shot object counting. *arXiv preprint arXiv:2305.07304*, 2023.

Yuhong Li, Xiaofan Zhang, and Deming Chen. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1091–1100, 2018.

Dingkang Liang, Wei Xu, and Xiang Bai. An end-to-end transformer model for crowd localization. In *European Conference on Computer Vision*, pp. 38–54, 2022.

Dingkang Liang, Jiahao Xie, Zhikang Zou, Xiaoqing Ye, Wei Xu, and Xiang Bai. CrowdCLIP: Unsupervised crowd counting via vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2893–2903, 2023.

Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, and Xiaopeng Hong. Boosting crowd counting via multi-faceted attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017.

Liang Liu, Hao Lu, Hongwei Xiong, Ke Xian, Zhiguo Cao, and Chunhua Shen. Counting objects by blockwise classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3513–3527, 2020.

Yiming Ma, Victor Sanchez, and Tanaya Guha. FusionCount: Efficient crowd counting via multiscale feature fusion. In *Proceedings of the IEEE International Conference on Image Processing*, pp. 3256–3260, 2022.

Yiming Ma, Victor Sanchez, and Tanaya Guha. CLIP-EBC: CLIP can count accurately through enhanced blockwise classification. *arXiv preprint arXiv:2403.09281*, 2025.

Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, 2021.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Oriane Siméoni et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3365–3374, 2021.

Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1595–1607, 2020.

Min Wang, Hao Cai, Yuchao Dai, and Ming Gong. Dynamic mixture of counter network for location-agnostic crowd counting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. NWPU-Crowd: A large-scale benchmark for crowd counting and localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2141–2149, 2021.

Haipeng Xiong, Hao Lu, Chengxin Liu, Liang Liu, Zhiguo Cao, and Chunhua Shen. From open set to closed set: Counting objects by spatial divide-and-conquer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8362–8371, 2019.

Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 589–597, 2016.

Kai Zhao, Chao He, Shuang Peng, and Tong Lu. A hybrid multi-scale transformer-cnn UNet for crowd counting. *Sensors*, 26(1):333, 2026.