Jurnal Elkolind Volume 13, Nomor 2, Juli 2026
DOI: http://dx.doi.org/10.33795/elkolind.v13i2.11416

Konfigurasi Bobot Supervisi Spasial Pada Penghitungan Kerumunan Berbasis Difusi untuk Dataset Dengan Mutu Anotasi Berbeda

Mas Nurul Achmadiah[1,2*)], Muhammad Ridha Agam[3]
[1]Department Electro-Optic Engineering, National Formosa University, Huwei Township, Yunlin County, Taiwan
[2]Jurusan Teknik Elektro, Politeknik Negeri Malang, Malang, Indonesia
[3]Department Electrical Engineering, National Formosa University, Huwei Township, Yunlin County, Taiwan
Penulis Korespondensi,* **e-mail: masnurul@polinema.ac.id

Received: 22/04/2026 *Revised: 25/07/2026* *Accepted: 25/07/2026*

ABSTRAK
Bobot komponen kerugian kerap dianggap sebagai rincian pelatihan yang sepele, padahal pada sistem penghitungan kerumunan yang dibangun di atas tulang punggung difusi beku hanya kepala kepadatan yang benar-benar dilatih, sehingga perimbangan antara kerugian relatif, kehalusan lokal, dan penyelarasan grid menentukan mutu peta kepadatan yang dihasilkan. Penelitian ini memetakan konfigurasi supervisi yang dibutuhkan lima dataset tolok ukur dengan tingkat kepadatan dan keterandalan anotasi berbeda, lalu menghubungkan tiap konfigurasi dengan sifat anotasi titiknya. Adegan yang sangat padat menuntut regularisasi struktural terkuat dengan bobot kehalusan lokal dan penyelarasan grid sebesar nol koma sepuluh, sedangkan dataset beranotasi tidak rapi menuntut penonaktifan kerugian relatif dan penyelarasan grid disertai laju pembelajaran terkecil. Dengan konfigurasi tersebut galat absolut rata-rata yang dicapai adalah 51,39 dan 5,79 pada dua bagian ShanghaiTech, 75,19 pada UCF-QNRF, serta 56,21 pada JHU-Crowd++, sehingga tetap rapat terhadap metode penghitungan terkini pada tiga dataset pertama namun tertinggal pada dataset terakhir. Temuan tersebut menghasilkan pedoman penyetelan yang menempatkan bobot supervisi berbasis skala dan supervisi tingkat grid sebagai turunan dari keterandalan anotasi titik, bukan semata dari tingkat kepadatan adegan.

Kata Kunci: bobot supervisi, fungsi kerugian, mutu anotasi, penghitungan kerumunan

ABSTRACT
Loss weighting is often treated as a minor training detail, yet in a crowd counting system built on a frozen diffusion backbone only the density head is actually trained, so the balance among the relative term, the local smoothness term, and the grid alignment term decides the quality of the predicted density map. This study maps the supervision configuration required by five benchmarks whose density level and annotation reliability differ, and relates each configuration to the properties of the point annotation. Very dense scenes require the strongest structural regularization, with local smoothness and grid alignment weighted at zero point ten, while a set with unreliable annotation requires the relative term and the grid alignment term to be switched off together with the smallest learning rate. Under those settings the mean absolute error reaches 51.39 and 5.79 on the two ShanghaiTech parts, 75.19 on UCF-QNRF, and 56.21 on JHU-Crowd++, which stays close to recent counting methods on the first three sets but falls behind on the last one. The finding yields a tuning rule in which the weight of scale based and grid level supervision follows the reliability of the point annotation rather than the density of the scene alone.

Keywords: *annotation quality, crowd counting, loss function, supervision weighting*

p-ISSN: 2356-0533; e-ISSN: 2355-9195

1. PENDAHULUAN

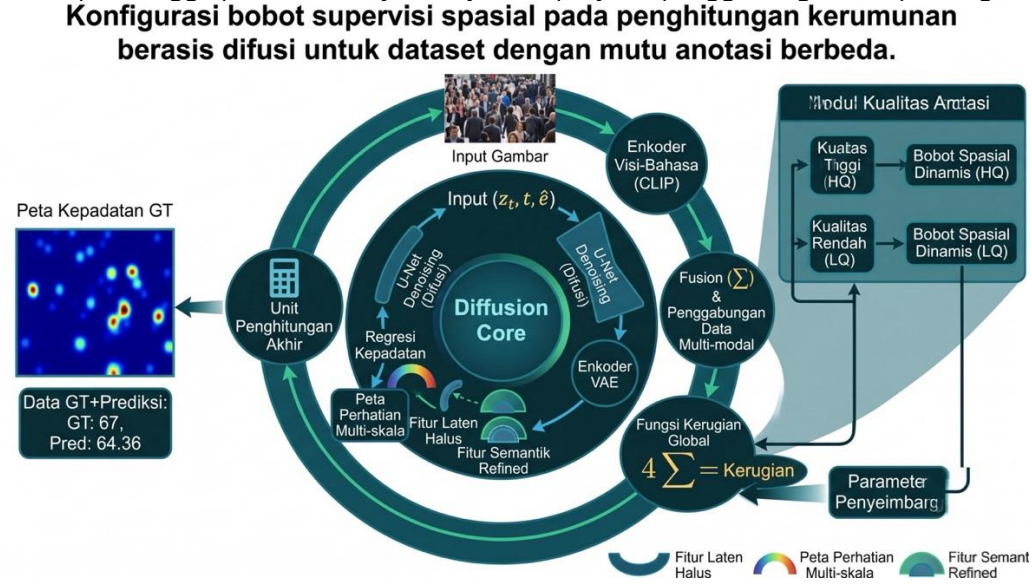
Penghitungan jumlah orang pada citra kerumunan dipakai luas untuk pemantauan ruang publik dan pengaturan arus penumpang [1]. Sebagian besar metode menempuh jalur regresi peta kepadatan, mulai dari jaringan berkolom banyak [2] dan konvolusi berdilatasi [3], hingga rancangan yang menekankan kecepatan inferensi melalui reparameterisasi struktural [4] serta rancangan yang menata ulang fungsi kerugian melalui pencocokan distribusi [5]. Perkembangan berikutnya menempatkan model praterlatih berskala besar sebagai tulang punggung yang dibekukan, baik model penyelaras citra dan teks [6] maupun model difusi, sehingga bagian yang benar-benar dilatih tinggal kepala regresi kepadatan beserta fungsi kerugiannya [7], [8] Pada susunan seperti itu mutu peta kepadatan tidak lagi ditentukan oleh kapasitas tulang punggung, melainkan oleh cara ketiga komponen supervisi disusun dan diberi bobot.

Nilai bobot tiap komponen kerugian umumnya dicantumkan tanpa penjelasan mengapa nilai tersebut dipilih, padahal setiap dataset tolok ukur memiliki tingkat kepadatan, rentang resolusi, dan ketelitian anotasi titik yang berlainan. Praktik menyalin bobot dari satu dataset ke dataset lain berisiko menghasilkan peta kepadatan yang terlalu rata pada adegan padat atau justru terlalu terikat pada anotasi yang tidak andal. Penelitian ini bertujuan memetakan konfigurasi laju pembelajaran dan bobot supervisi yang diperlukan lima dataset dengan sifat anotasi berbeda, menghubungkan tiap konfigurasi dengan ciri anotasinya, kemudian menilai galat yang dicapai terhadap sejumlah metode penghitungan yang lazim dipakai sebagai pembanding. Hasilnya dirangkum menjadi pedoman penyetelan sederhana yang dapat dipakai ketika model dipindahkan ke dataset atau lokasi pemasangan baru.

2. METODE PENELITIAN

2.1 Ringkasan sistem dan susunan supervisi

Sistem bekerja dalam dua alur sebagaimana Gambar 1. Alur pelatihan mengambil fitur visual dan fitur tekstual dari encoder praterlatih, menyatukannya melalui proyeksi ringan, lalu memakai gabungan tersebut sebagai konteks cross-attention pada U-Net difusi yang dibekukan. Alur penghitungan menerima citra beserta laten berderau, langkah waktu, dan injeksi derau, kemudian menghasilkan fitur laten dan fitur semantik lintas skala yang dipakai oleh kepala regresi kepadatan. Seluruh bobot encoder, VAE, dan U-Net dipertahankan tetap, sehingga pelatihan hanya menyentuh proyeksi penggabung dan kepala regresi.



Gambar 1: Susunan sistem penghitungan kerumunan berbasis difusi beserta letak ketiga komponen supervisi pada alur pelatihan

Bagian kiri Gambar 1 memperlihatkan alur masukan yang membentuk konteks bagi U-Net difusi: citra kerumunan diteruskan sekaligus ke encoder visi-bahasa CLIP dan ke encoder VAE, sehingga diperoleh

laten awal z_t beserta suntikan langkah waktu t dan embedding teks ê sebagai masukan Diffusion Core. Selama proses denoising, U-Net difusi menghasilkan tiga keluaran yang disatukan pada tahap fusi dan penggabungan data multi-modal, yaitu peta perhatian multi-skala, fitur laten yang telah dihaluskan, dan fitur semantik yang telah disaring (refined). Ketiga keluaran inilah yang menjadi fitur laten pada persamaan (1) dan yang diregresikan oleh kepala kepadatan menjadi peta D, sebelum unit penghitungan akhir menjumlahkan nilai piksel peta tersebut menjadi estimasi C.

Bagian kanan Gambar 1 menggambarkan mekanisme penentuan bobot yang dibahas pada subbagian 2.2, yaitu modul kualitas anotasi yang mengelompokkan dataset menjadi kualitas tinggi (HQ) dan kualitas rendah (LQ), lalu memetakan tiap kelompok ke bobot spasial dinamis yang berbeda. Kelompok HQ diarahkan ke bobot spasial yang lebih besar sebagaimana UCF-CC-50 pada Tabel I, sedangkan kelompok LQ diarahkan ke bobot yang diperkecil atau dinonaktifkan sebagaimana JHU-Crowd++. Bobot spasial dinamis ini menjadi parameter penyeimbang yang disuntikkan ke fungsi kerugian global, yang pada Gambar 1 dituliskan sebagai penjumlahan empat komponen (4Σ = Kerugian): kerugian kepadatan tingkat piksel, kerugian relatif di dalam Lglobal, kerugian kehalusan lokal, dan kerugian penyelarasan grid, dua komponen terakhir yang pada persamaan (2) dirangkum sebagai Lstruct. Dengan susunan tersebut, contoh pasangan Data GT+Prediksi pada Gambar 1 (GT 67, prediksi 64,36) menggambarkan selisih hitungan pada satu citra yang kemudian diringkas secara agregat melalui MAE dan RMSE pada persamaan (3) dan (4).

Rumusan prompt yang dipakai dijaga tetap pada seluruh percobaan agar konfigurasi supervisi menjadi satu-satunya peubah yang diamati. Dengan cara itu perbedaan galat antar dataset dapat dibaca sebagai akibat pemilihan bobot dan laju pembelajaran, bukan akibat perbedaan konteks tekstual yang masuk ke lapisan atensi.

Peta kepadatan dihasilkan melalui regresi konvolusi ringan atas fitur laten yang telah disatukan dengan grid semantik, sedangkan jumlah orang diperoleh dengan menjumlahkan seluruh nilai piksel peta tersebut seperti pada (1).

$$C = sigma\_(x, y)\, D(x, y) \tag{1}$$

dengan D sebagai peta kepadatan prediksi dan C sebagai jumlah orang hasil estimasi. Fungsi kerugian pelatihan tersusun atas tiga komponen seperti pada (2).

$$\mathrm{L_{total}} = \mathrm{L_{dens}} + \mathrm{L_{global}} + \mathrm{L_{sreuct}} \tag{2}$$

dengan Ldens sebagai kerugian regresi kepadatan tingkat piksel, Lglobal sebagai kerugian penghitungan global yang di dalamnya terdapat komponen kerugian relatif, dan Lstruct sebagai gabungan regularisasi kehalusan lokal dan penyelarasan tingkat grid. Kerugian relatif menormalkan selisih hitungan terhadap jumlah acuan sehingga citra bermuatan sedikit orang tidak tenggelam oleh citra bermuatan ribuan orang. Kehalusan lokal menekan lonjakan nilai antar piksel bertetangga, sedangkan penyelarasan grid menuntut kecocokan jumlah pada tiap petak citra dan bukan hanya pada citra utuh. Ketiga komponen tersebut sama-sama bersandar pada ketepatan posisi anotasi titik, sehingga bobotnya perlu ditimbang menurut mutu anotasi tiap dataset.

2.2 Dataset dan prosedur penyetelan

Lima dataset dipakai dengan ciri yang sengaja dibuat berlainan. ShanghaiTech Part A memuat adegan padat dengan sudut pandang beragam, sedangkan Part B memuat adegan jalan dengan kepadatan rendah dan resolusi seragam. UCF-CC-50 memiliki jumlah citra sedikit dengan tingkat kepadatan sangat tinggi. UCF-QNRF mencakup rentang resolusi dan kepadatan paling lebar di antara kelima dataset. JHU-Crowd++ memuat oklusi berat serta anotasi titik yang tidak selalu tepat posisi.

Penyetelan dilakukan per dataset dengan menyesuaikan laju pembelajaran dan tiga bobot supervisi, yaitu bobot kerugian relatif, bobot kehalusan lokal, dan bobot penyelarasan grid. Nilai yang dilaporkan adalah nilai yang memberi galat terkecil pada data uji tiap dataset. Seluruh percobaan dijalankan pada GPU NVIDIA RTX 5090 dengan encoder dan U-Net difusi dalam keadaan beku.

2.3 Metrik evaluasi

Ketelitian dinilai dengan galat absolut rata-rata seperti pada (3) dan akar galat kuadrat rata-rata seperti pada (4) yang keduanya lazim dipakai pada penelitian penghitungan kerumunan.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |C_i - C_{i_{acuan}}| \quad (3)$$

$$\mathrm{RMSE} = \left(\sqrt{\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} |\mathrm{C_i} - \mathrm{C_{i_{acuan}}}|} \right)^2 \quad (4)$$

dengan N sebagai jumlah citra uji, Ci sebagai jumlah orang hasil estimasi, dan Ciacuan sebagai jumlah orang menurut anotasi. Galat absolut rata-rata dipakai sebagai tolok ukur utama karena mencerminkan besar kesalahan hitungan rata-rata, sedangkan akar galat kuadrat rata-rata dipakai untuk menyoroti pengaruh kesalahan besar yang biasanya muncul pada citra berkepadatan tinggi.

3. HASIL DAN PEMBAHASAN

3.1 Konfigurasi supervisi yang dibutuhkan tiap dataset

Tabel I memuat laju pembelajaran dan bobot ketiga komponen supervisi yang dipakai pada tiap dataset.

TABEL I : KONFIGURASI LAJU PEMBELAJARAN DAN BOBOT SUPERVISI TIAP DATASET

Dataset	Laju pembelajaran	Bobot kerugian relatif	Bobot kehalusan lokal	Bobot penyelarasan grid
ShanghaiTech Part A	5×10^{-6}	0,03	0,05	0,05
ShanghaiTech Part B	5×10^{-6}	0,02	0,02	0,03
UCF-CC-50	3×10^{-6}	0,05	0,10	0,10
UCF-QNRF	1×10^{-5}	0,04	0,05	0,05
JHU-Crowd++	1×10^{-6}	0,00	0,02	0,00

Pola pada Tabel I terbagi menjadi tiga kelompok. Dataset berkepadatan sangat tinggi seperti UCF-CC-50 menuntut regularisasi struktural terkuat, yaitu 0,10 pada kehalusan lokal maupun penyelarasan grid, karena tumpukan kepala yang rapat membuat peta kepadatan mudah pecah menjadi puncak sempit yang terpisah. Dataset berkepadatan rendah seperti ShanghaiTech Part B memerlukan bobot terkecil di antara dataset yang masih mengaktifkan seluruh komponen, yaitu 0,02 pada kerugian relatif dan kehalusan lokal serta 0,03 pada penyelarasan grid, sebab adegan yang lengang membuat regularisasi berlebih meratakan puncak yang seharusnya tajam. JHU-Crowd++ berada di luar kedua pola tersebut karena kerugian relatif dan penyelarasan grid dimatikan sepenuhnya dan laju pembelajaran ditekan hingga 1×10^{-6}. Pilihan itu diperlukan karena oklusi berat dan anotasi titik yang bergeser membuat supervisi berbasis skala dan supervisi tingkat petak menyampaikan isyarat yang keliru. UCF-QNRF menempati posisi tengah dengan laju pembelajaran terbesar, yaitu 1×10^{-5}, yang sejalan dengan lebarnya ragam resolusi sehingga kepala regresi memerlukan langkah pembaruan lebih besar untuk mencakup seluruh skala.

3.2 Galat yang dicapai dan posisinya terhadap metode pembanding

Tabel II menempatkan galat yang dicapai dengan konfigurasi pada Tabel I berdampingan dengan sejumlah metode penghitungan yang lazim dipakai sebagai pembanding.

TABEL II : PERBANDINGAN GALAT ABSOLUT RATA-RATA PADA EMPAT DATASET TOLOK UKUR

Metode	SH Part A	SH Part B	UCF-QNRF	JHU-Crowd++
CSRNet	68,2	10,6	120,3	-
DM-Count	59,7	7,4	85,6	-
DSGC-Net	48,9	5,9	79,3	-
LIMM	50,8	6,5	76,4	53,0
CLIP-EBC	52,5	6,6	80,3	-
Model yang diuji	51,39	5,79	75,19	56,21

Pada ShanghaiTech Part B model mencatat galat 5,79 yang merupakan nilai terendah pada tabel, sedangkan pada UCF-QNRF nilai 75,19 juga lebih rendah daripada seluruh pembanding. Pada ShanghaiTech Part A nilai 51,39 masih berada di atas DSGC-Net yang mencatat 48,9 [9] namun tetap lebih rendah daripada CLIP-EBC [7] yang mencatat 52,5. Hasil pada JHU-Crowd++ menunjukkan arah sebaliknya, yaitu 56,21 berbanding 53,0 milik LIMM. Selisih tersebut sejalan dengan konfigurasi pada Tabel I. Pada dataset itu kerugian relatif dan penyelarasan grid dimatikan, sehingga kepala regresi hanya bersandar pada kerugian kepadatan tingkat piksel dan kehalusan lokal berbobot 0,02. Penonaktifan yang diperlukan untuk menghindari anotasi yang tidak andal karenanya dibayar dengan ketelitian yang menurun, dan besar biaya tersebut terbaca langsung dari selisih galat.

3.3 Pedoman penyetelan dan implikasi penerapan

Tiga aturan sederhana dapat ditarik dari kedua tabel. Pertama, bobot kehalusan lokal dan penyelarasan grid dinaikkan mengikuti tingkat kepadatan adegan, dengan 0,10 sebagai batas atas yang terpakai pada dataset terpadat. Kedua, komponen yang bersandar pada posisi titik, yaitu kerugian relatif dan penyelarasan grid, diturunkan atau dimatikan ketika anotasi diperkirakan bergeser, tanpa memandang seberapa padat adegannya. Ketiga, laju pembelajaran diperbesar ketika ragam resolusi melebar dan diperkecil ketika supervisi struktural dikurangi, agar pelatihan tidak bertumpu terlalu berat pada satu komponen kerugian yang tersisa.

Karena bagian yang dilatih hanya proyeksi penggabung dan kepala regresi, penyetelan ulang pada lokasi baru berlangsung murah dan tidak menuntut penggantian tulang punggung. Sifat tersebut sesuai dengan kebutuhan sistem tertanam yang mengutamakan pemeliharaan sederhana dan konsumsi daya rendah [10], [11] Kebutuhan serupa juga muncul pada penerapan visi komputer lain seperti kendali kendaraan tanpa awak [12] klasifikasi citra bermotif [13] dan penghitungan agnostik kelas [14].

3.4 Keterbatasan dan arah pengembangan

Konfigurasi pada Tabel I diperoleh melalui penelusuran manual, sehingga tidak ada jaminan bahwa nilai tersebut merupakan titik terbaik pada ruang pencarian. Penelitian ini juga belum mengukur seberapa peka galat terhadap perubahan kecil pada tiap bobot, sehingga lebar rentang aman untuk setiap komponen masih belum diketahui.

Pengembangan berikutnya dapat menempuh dua arah. Arah pertama adalah penyetelan bobot secara otomatis berdasarkan ukuran keterandalan anotasi yang dihitung langsung dari data, sehingga penonaktifan komponen tidak lagi ditetapkan secara manual. Arah kedua adalah pengujian pada data lapangan dengan ciri anotasi yang berbeda dari kelima dataset tolok ukur, termasuk data yang dianotasi oleh lebih dari satu penganotasi.

4. KESIMPULAN

Penelitian ini memetakan kebutuhan konfigurasi supervisi pada penghitungan kerumunan berbasis difusi untuk lima dataset dengan ciri anotasi berbeda. Dataset berkepadatan sangat tinggi seperti UCF-CC-50 memerlukan bobot kehalusan lokal dan penyelarasan grid sebesar 0,10, dataset berkepadatan rendah seperti ShanghaiTech Part B cukup dengan bobot 0,02 hingga 0,03, sedangkan JHU-Crowd++ menuntut penonaktifan kerugian relatif dan penyelarasan grid disertai laju pembelajaran 1×10^{-6}. Dengan konfigurasi tersebut model mencatat galat absolut rata-rata 51,39 pada ShanghaiTech Part A, 5,79 pada Part B, 75,19

pada UCF-QNRF, dan 56,21 pada JHU-Crowd++. Perbandingan itu memperlihatkan bahwa bobot supervisi sebaiknya mengikuti keterandalan anotasi titik dan bukan semata tingkat kepadatan adegan. Pengujian pada data lapangan dan penyetelan bobot secara otomatis menjadi langkah lanjutan yang direncanakan.

DAFTAR PUSTAKA

[1] M. N. Achmadiah, N. Setyawan, A. A. Bryantono, C.-C. Sun, and W.-K. Kuo, "Fast Person Detection Using YOLOX With AI Accelerator For Train Station Safety," in *2024 International Electronics Symposium (IES)*, IEEE, Aug. 2024, pp. 504–509. doi: 10.1109/IES63037.2024.10665874.

[2] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 589–597. doi: 10.1109/CVPR.2016.70.

[3] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 1091–1100. doi: 10.1109/CVPR.2018.00120.

[4] M. N. Achmadiah, C.-C. Sun, W.-K. Kuo, and J.-W. Hsieh, "RepSFNet : A Single Fusion Network with Structural Reparameterization for Crowd Counting," in *2025 IEEE International Conference on Advanced Visual and Signal-Based Systems (AVSS)*, IEEE, Aug. 2025, pp. 1–6. doi: 10.1109/AVSS65446.2025.11149882.

[5] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai, "Distribution Matching for Crowd Counting," in *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.*, 2020.

[6] Alec Radford * 1 Jong Wook Kim * 1 Chris Hallacy 1 Aditya Ramesh 1 Gabriel Goh 1 Sandhini Agarwal 1 Girish Sastry 1 Amanda Askell 1 Pamela Mishkin 1 Jack Clark 1 Gretchen Krueger 1 Ilya Sutskever 1, "Learning Transferable Visual Models From Natural Language Supervision," in *38 th International Conference on Machine Learning*, 2021.

[7] Y. Ma, V. Sanchez, and T. Guha, "CLIP-EBC: CLIP Can Count Accurately through Enhanced Blockwise Classification," in *2025 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, Jun. 2025, pp. 1–6. doi: 10.1109/ICME59968.2025.11209839.

[8] N. Amini-Naieni, T. Han, and A. Zisserman, "CountGD: Multi-Modal Open-World Counting," in *Advances in Neural Information Processing Systems 37*, San Diego, California, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024, pp. 48810–48837. doi: 10.52202/079017-1547.

[9] Yihong wu, jinqiau wei, xionghui zhao, and yidi li, "DSGC-Net: A Dual-Stream Graph Convolutional Network for Crowd Counting via Feature Correlation Mining,"

[10] N. Setyawan, C.-C. Sun, M.-H. Hsu, W.-K. Kuo, and J.-W. Hsieh, "MicroViT: A Vision Transformer with Low Complexity Self Attention for Edge Device," in *2025 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, May 2025, pp. 1–5. doi: 10.1109/ISCAS56072.2025.11043206.

[11] M. Nurul Achmadiah, A. Ahamad, C.-C. Sun, and W.-K. Kuo, "Energy-Efficient Fast Object Detection on Edge Devices for IoT Systems," *IEEE Internet Things J.*, vol. 12, no. 11, pp. 16681–16694, Jun. 2025, doi: 10.1109/JIOT.2025.3536526.

[12] G. Al Azhar, S. Sungkono, M. N. Achmadiyah, and S. Izza, "Peningkatan Kestabilan Sistem Kontrol UGV melalui Optimalisasi Manajemen Core dan Free-RTOS pada ESP32," *Jurnal Elektronika dan Otomasi Industri*, vol. 10, no. 2, pp. 253–263, Jul. 2023, doi: 10.33795/elkolind.v10i2.3720.

[13] N. Setyawan, M. N. Achmadiah, C.-C. Sun, and W.-K. Kuo, "Multi-Stage Vision Transformer for Batik Classification," in *2024 International Electronics Symposium (IES)*, IEEE, Aug. 2024, pp. 449–453. doi: 10.1109/IES63037.2024.10665807.

[14] Q. W. * , H. R. and J. L. Xiaofei Hui, "Class-Agnostic Object Counting with Text-to-Image Diffusion Model," in *European Computer Vision Association (ECVA)*,