

Estimasi Jumlah dan Kepadatan Orang Pada Citra Kerumunan Tunggal Menggunakan Metode Convolutional Neural Network (CNN)

Deu Aldo Dhavicky, Reza Fuad Rachmadi, Supeno Mardi Susiki Nugroho

Dept. Teknik Komputer Institut Teknologi Sepuluh Nopember Surabaya, Indonesia

Email: deu15@mhs.te.its.ac.id, fuad@its.ac.id, mardi@te.its.ac.id

Abstrak

Estimasi jumlah orang pada pusat keramaian saat ini banyak diterapkan, baik menggunakan cara konvensional seperti menghitung secara manual hingga menggunakan bantuan alat atau sensor, hal ini dilakukan dengan tujuan untuk memantau jumlah populasi, karena populasi orang dengan jumlah besar yang berkumpul pada suatu titik di area tertentu akan memunculkan berbagai masalah, salah satunya keamanan. Salah satu metode yang saat ini sedang dikembangkan untuk mendeteksi dan mengenali object pada sebuah citra dengan menggunakan (*Convolutional Neural Network*) (CNN). CNN adalah pengembangan dari *Multilayer Perceptron* (MLP) yang dirancang untuk mengolah data dua dimensi. Dalam tugas akhir ini dikembangkan estimasi jumlah dan kepadatan orang pada citra kerumunan tunggal menggunakan metode *Convolutional Neural Network* dengan menggunakan Shanghai Tech datasets.

Kata kunci: Crowd, Crowd-Counting, Klasifikasi, Convolutional Neural Network (CNN)

Diterima Redaksi: 05-Juni-2025 Selesai Revisi: 14-Juni-2025 Diterbitkan Online: 15-Juli-2025
DOI: <https://doi.org/10.59378/jcenim.v3i2.74>

I. PENDAHULUAN

Pada tempat – tempat yang terdapat banyak kerumunan orang (seperti stadion, alun-alun kota, pusat pameran dan sebagainya), terutama di pintu masuk atau persimpangan merupakan fenomena yang belakangan cukup membutuhkan perhatian. Terdapat populasi orang dengan jumlah besar yang berkumpul pada suatu titik di area tertentu akan memunculkan berbagai masalah, salah satunya keamanan. Fenomena tersebut memunculkan risiko adanya tindak kejahatan hingga terjadinya kecelakaan yang dapat mengakibatkan cedera serius atau berakibat kematian. Oleh karena itu dibutuhkan manajemen informasi yang baik mengenai fenomena jumlah orang dan kepadatan kerumunan yang ada di suatu tempat [1].

Sistem penghitungan orang (*people counting*) adalah sebuah sistem yang mampu menghitung orang berdasarkan citra yang ditangkap, dengan tingkat pencahayaan yang cukup [2]. Masalah penghitungan orang ini sebenarnya dapat diselesaikan dengan cara konvensional seperti penghitungan secara manual atau menggunakan sensor, namun hal ini memiliki kelemahan, selain memakai sumber daya manusia, cara tersebut juga membuat tidak nyaman dan membatasi ruang gerak. Sejumlah pendekatan telah diusulkan untuk mengatasi perhitungan kerumunan dan mereka dapat diklasifikasikan menjadi tiga jenis metode, berbasis deteksi objek, berbasis regresi global dan berbasis estimasi kepadatan Metode berbasis deteksi objek [3] selalu menggunakan detektor objek visual, yang akan menangkap objek individu yang terdapat dalam gambar kemudian menjumlahkan objek tersebut. Tetapi, sulit untuk mendapatkan hasil objek individu saat keadaan ramai dengan kondisi hasil tangkap citra kurang baik, ukuran kecil atau adanya blur. Karena metode berbasis deteksi objek kurang dapat diaplikasikan pada keadaan dengan situasi tertentu, sejumlah metode lain mencoba memperkirakan jumlah kerumunan dengan regresi dengan fitur global. Metode regresi termasuk *ridge regression*, *Neural network* dan beberapa lainnya. Dibandingkan dengan metode berbasis deteksi, penghitungan secara regresi global dapat lebih di aplikasikan dalam situasi yang lebih sulit. Selanjutnya metode berbasis estimasi kepadatan, yang dapat mempertahankan lebih banyak informasi spasial dari citra yang di dapat [4]. Oleh karena itu, pada tugas akhir ini diusulkan sebuah pengembangan terhadap estimasi jumlah dan kepadatan pada kerumunan menggunakan metode *Convolutional Neural Network* (CNN).

II. RELATED WORK

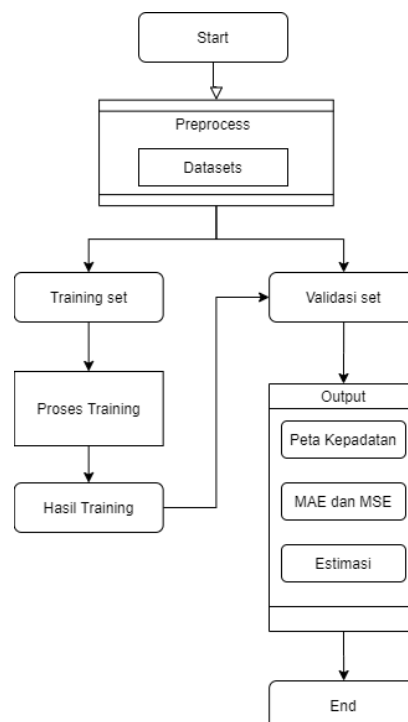
A. Counting Crowd with Fully Convolutional Networks

Penelitian berjudul *Counting Crowd with Fully Convolutional Networks* yang dilakukan oleh Jianyong Wang, Lu Wang, dan Fenglei Yang ini mengusulkan pendekatan yang dipelajari secara menyeluruh dari patch gambar dengan merevisi distribusi kepadatan kerumunan. Model kerumunan FCN ini dapat menampilkan peta kepadatan kerumunan tinggi dan kuantitas kerumunan dapat diintegrasikan oleh peta. Selain itu, untuk menangani masalah distorsi perspektif adegan, kami mengusulkan metode generasi kebenaran kepadatan tanah. Hasil percobaan menunjukkan bahwa metode penghitungan kerumunan kami mencapai akurasi terbaik dengan menggunakan dataset WorldExpo 10 dibandingkan dengan metode lainnya. [4]

B. Single-Image Crowd Counting via Multi Column Convolutional Neural Network

Penelitian berjudul *Single-Image Crowd Counting via Multi-Column Convolutional Neural Network* [5] yang dilakukan oleh Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, dan Yi Ma ini merupakan pendekatan yang sederhana namun efektif untuk memetakan gambar ke peta keramaian keramaian. MCNN yang diusulkan memungkinkan gambar input berukuran atau resolusi yang berubah-ubah. Dengan menggunakan filter dengan bidang reseptif dari berbagai ukuran, fitur yang dipelajari oleh setiap kolom CNN adaptif terhadap variasi dalam ukuran orang / kepala karena efek perspektif atau resolusi gambar. Selain itu, peta kepadatan sebenarnya dihitung secara akurat berdasarkan kernel geometri adaptif yang tidak perlu mengetahui peta perspektif gambar input. Karena kumpulan data penghitungan kerumunan yang ada tidak cukup untuk mencakup semua situasi yang menantang, telah mengumpulkan dan memberi label dataset baru yang besar yang mencakup 1.198 gambar dengan sekitar 330.000 kepala. Pada dataset baru ini, beserta semua dataset yang ada, dilakukan percobaan ekstensif untuk memverifikasi efektivitas model dan metode yang diusulkan. Secara khusus, dengan model MCNN sederhana yang diusulkan. Selain itu, percobaan menunjukkan bahwa model yang disebut, setelah dilatih pada satu dataset, dapat dengan mudah ditransfer ke dataset baru.

III. DESAIN DAN IMPLEMENTASI SISTEM



Gambar 1: Diagram alir

Tujuan dari penelitian ini adalah untuk melakukan pengembangan terhadap estimasi jumlah dan memberikan peta kepadatan orang di suatu tempat berbasis citra tunggal dengan menggunakan metode *Convolutional Neural Network*. Proses kerja pada penelitian ini ditunjukkan pada Gambar 1. Berdasarkan pada Gambar 1, salah satu proses yang dilakukan adalah melakukan sebuah pendekatan pada peta kepadatan. Dataset di berikan koordinat perkiraan posisi tengah setiap kepala manusia di setiap gambar kerumunan yang telah diberi label secara manual. Agar bisa mencapai hal tersebut. Dataset yang digunakan untuk *training* dalam sistem ini digunakan *Kernel Gaussian* pada saat melakukan *preprocess* dataset. Sehingga didapatkan wilayah yang diubah menjadi probabilitas bahwa wilayah tersebut merupakan kepala manusia. Setelah mendapatkan dataset, langkah selanjutnya adalah pembuatan *split* yang berfungsi untuk memisahkan dataset yang digunakan untuk *training* dan *testing* atau validasi. Metode yang digunakan pada sistem ini adalah CNN yang menggunakan arsitektur berdasarkan pada model *Multi-Column Convolutional Neural Network* dan *Cascaded Multi-task Learning*. Hasil dari *training* tersebut adalah weight yang berupa model terbaik dalam proses *training* sebagai *validation set*. Proses selanjutnya adalah validasi model, yaitu proses dimana validasi model yang telah di proses dari dataset yang digunakan yang akan menghasilkan peta kepadatan.

A. Desain Sistem

Secara garis besar, terdapat empat proses yang ada pada bagian ini.

1. Dataset

Dataset yang digunakan pada tugas akhir ini adalah dataset ShanghaiTech yang berisi 1198 gambar yang sudah terlabeli, dengan total 330.165 objek orang pada gambar yang terlabeli. Serta menggunakan 14 gambar yang diambil dari internet dengan latar tempat Kota Surabaya.

2. Preprocessing Data

Sebelum digunakan, dataset perlu diproses terlebih dahulu. Dataset yang akan digunakan akan diubah menjadi 9 bagian set gambar, 9 sub-bagian tersebut memiliki masing-masing ukuran 1/4 dari ukuran gambar aslinya. Kemudian dilakukan *split* untuk memisahkan data yang akan digunakan sebagai input *training* dan validasi.

3. Training

Training merupakan sebuah proses pelatihan model CNN menggunakan dataset ShanghaiTech. *Training* pada sistem ini akan dilakukan dengan menggunakan model dari *Multi-Column Convolutional Neural Network* dan *Cascaded-MtL* atau *Cascaded Multi-task Learning*.

4. Evaluasi

Model yang telah melewati proses *training* kemudian akan dievaluasi untuk menilai performanya. Evaluasi model akan melakukan pengujian terhadap model hasil *training*.

B. Dataset

Pada penelitian ini digunakan dataset Shanghai tech yang berisi 2 jenis bagian, yaitu Part A dan Part B. Pada bagian Part A dataset tersebut 482 gambar yang terdapat didalamnya merupakan kumpulan gambar yang didapat dari Internet dan dengan menggunakan keyword pencarian yang bervariasi. Sedangkan pada bagian Part B dataset tersebut terdapat sebanyak 716 gambar yang terdapat didalamnya yang didapat dari jalanan metropolitan pada daerah Shanghai. Kedua data tersebut baik Part A dan Part B kemudian dibedakan lagi untuk *training* dan *testing*: 300 gambar pada Part A digunakan untuk *training* dan 182 gambar untuk *testing*; serta 400 gambar pada Part B untuk *training* dan 316 gambar untuk *testing*.

Tabel 1: ShanghaiTech

Fitur	Part A	Part B
Resolusi berbeda	768 x 1024	
Jumlah	482	716
Label total	241.667	88.488

1. Mengolah Dataset

Karena dataset yang digunakan adalah gambar, dataset yang telah didapat diubah dengan menggunakan program yang sudah dibuat. Dataset diubah menjadi 9 bagian set gambar, di mana 9 sub-bagian tersebut masing-masing memiliki ukuran $1/4$ dari ukuran gambar aslinya.



Gambar 2: Processing dataset

2. Split Dataset

Dataset ShanghaiTech terdiri dari dua bagian, yaitu Part A dan Part B:

- **Part A:** Berisi 482 gambar dari internet (300 gambar untuk *training* dan 182 gambar untuk *testing*).
- **Part B:** Berisi 716 gambar dari jalanan metropolitan Shanghai (400 gambar untuk *training* dan 316 gambar untuk *testing*).

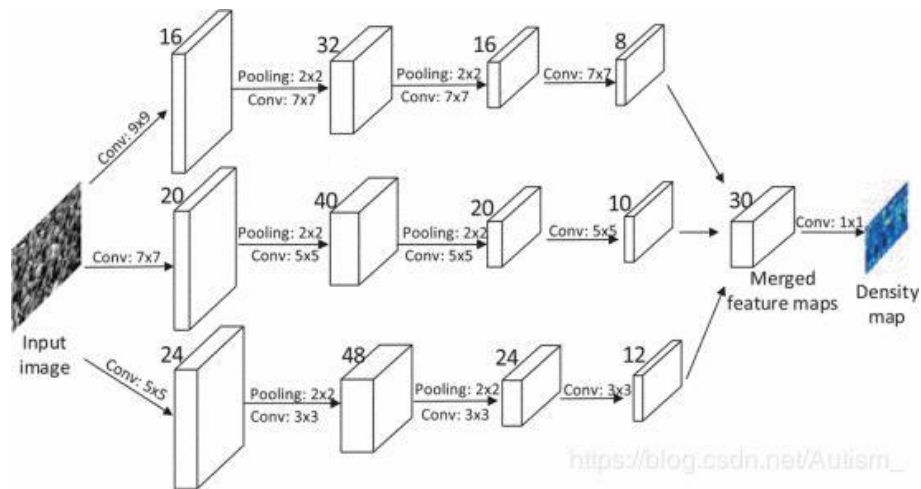
3. Proses Training

Pada proses ini, data *train* yang telah diubah menjadi *sub-block* (ukuran $1/4$ asli) digunakan untuk membentuk pola berupa bobot (*weight*). *Training* dilakukan menggunakan model *Multi-Column Convolutional Neural Networks* (MCNN) dan *Cascaded Multi-task Learning* (*Cascaded-MtL*). Dalam proses ini, ditentukan dua parameter utama:

- Learning Rate:** Parameter kontrol untuk memperbarui bobot model berdasarkan estimasi *error*.
- Epoch:** Satu set putaran penuh pelatihan. Pada sistem ini, jumlah *epoch* ditetapkan sebanyak 2000 kali untuk meminimalkan nilai *loss*.

C. Training Models

Pada penelitian ini, digunakan model MCNN dan *Cascaded-MtL*. MCNN yang diusulkan memungkinkan gambar input berukuran atau resolusi yang berubah-ubah. Dengan menggunakan filter dengan bidang reseptif dari berbagai ukuran, fitur yang dipelajari oleh setiap kolom CNN adaptif terhadap variasi dalam ukuran orang / kepala karena efek perspektif atau resolusi gambar. Selain itu, peta kerapatan sebenarnya dihitung secara akurat berdasarkan kernel geometri-adaptif yang tidak perlu mengetahui peta perspektif gambar input. Sedangkan *Cascaded-MtL* adalah untuk mempelajari model-model yang memenuhi berbagai tingkat kepadatan yang disajikan dengan menggabungkan *high-level prior* sebelum masuk ke network. *high-level prior* di tuju untuk mengklasifikasikan penghitungan ke dalam berbagai kelompok yang label kelasnya didasarkan pada jumlah orang yang hadir dalam gambar. Dengan mengeksplorasi jumlah label, metode ini memiliki kemampuan untuk menghitung jumlah orang di dalam peta, terlepas dari berbagai macam skala yang digunakan dalam jaringan.



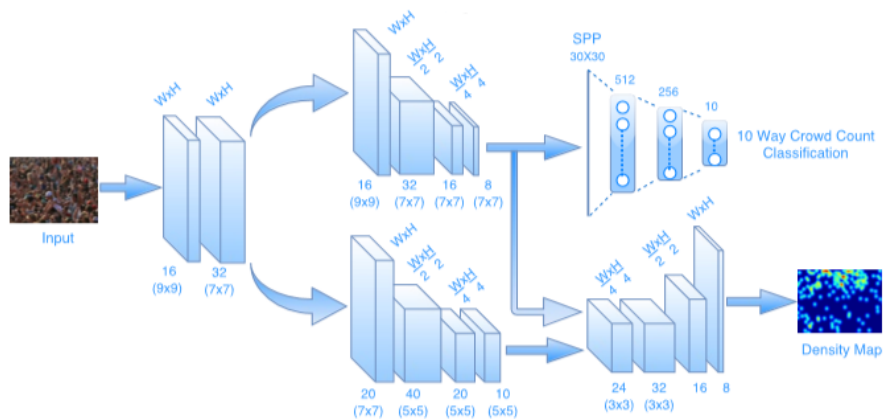
Gambar 3: Susunan arsitektur pada model pertama

1. Model Pertama (MCNN)

Model pertama menggunakan arsitektur *Multi-Column Convolutional Neural Network* (MCNN). Sesuai pada Gambar 4, model ini terdiri dari tiga kolom paralel:

- **L-column** (Kernel besar): Konfigurasi ukuran 9×9, 7×7, 7×7, dan 7×7.
- **M-column** (Kernel sedang): Konfigurasi ukuran 7×7, 5×5, 5×5, dan 5×5.
- **S-column** (Kernel kecil): Konfigurasi ukuran 5×5, 3×3, 3×3, dan 7×7.

Setiap kolom menggunakan 2D *Convolution Layer* yang diikuti dengan 2D *max pooling layer* berukuran 2×2 pada layer 1 dan layer 2. Model ini bersifat *fully-connected* dan menggunakan fungsi aktivasi ReLU.



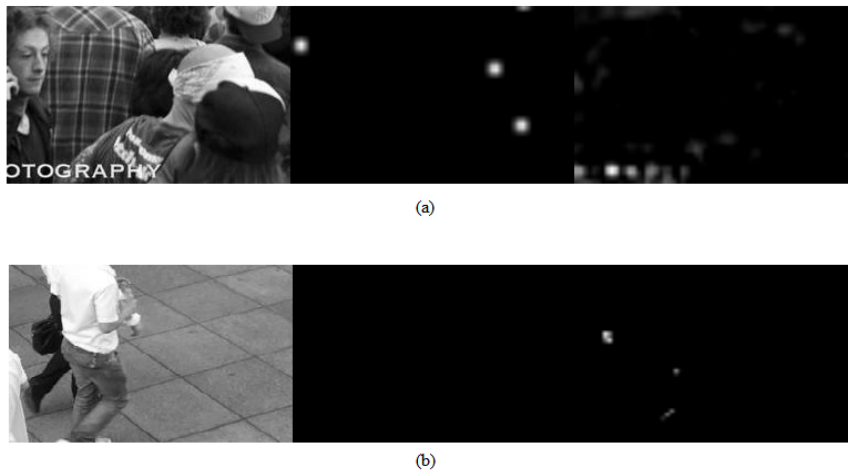
Gambar 4: Susunan arsitektur pada model pertama

2. Model Kedua (Cascaded-MtL)

Model kedua terdiri dari dua tahap utama yang berbagi set fitur konvolusional yang sama:

- **Tahap Pertama (High-level Prior):** Terdiri dari 2 lapisan konvolusional dengan aktivasi PReLU. Layer pertama memiliki 16 peta fitur (9×9) dan layer kedua menggunakan filter 7×7.
- **Tahap Kedua (Estimasi Peta Kepadatan):** Terdiri dari 4 lapisan konvolusional dengan aktivasi PReLU. Dua lapisan pertama menggunakan *max pooling* (*stride* 2). Konfigurasi filternya adalah: 20 peta fitur (7×7), 40 peta fitur (5×5), 20 peta fitur (5×5), dan 10 peta fitur (5×5).

Output dari jaringan estimasi dikombinasikan dengan lapisan terakhir dari *high-level prior* melalui dua konvolusi penutup dengan ukuran filter 3×3 (24 dan 32 peta fitur).



Gambar 5: Susunan arsitektur pada model kedua

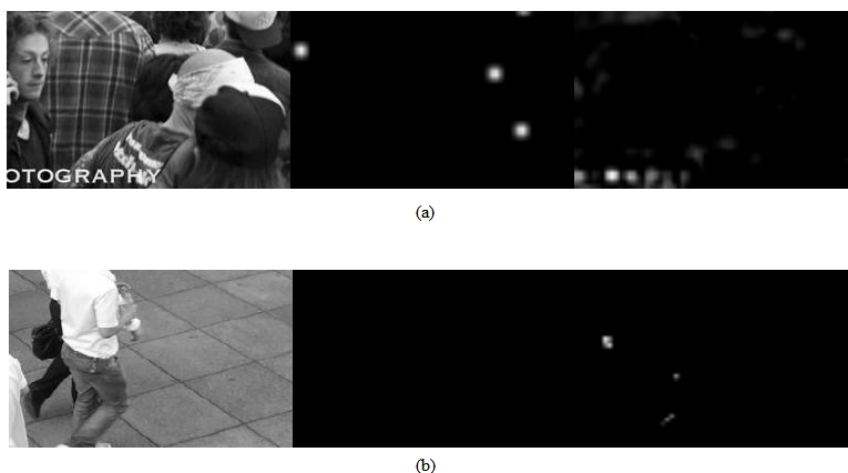
D. Evaluasi

Setelah proses *training* selesai, proses selanjutnya adalah melakukan evaluasi terhadap hasil *training* tersebut. Evaluasi ini bertujuan untuk melakukan pengecekan terhadap proses *training* model yang telah dilakukan. Untuk evaluasi model digunakan metode dengan *Mean Absolute Error* (MAE) dan *Mean Square Error* (MSE).

IV. PENGUJIAN DAN ANALISA

A. Training Data

Proses *training* data dilakukan sebanyak 10 kali dengan menggunakan *Convolutional Neural Network* dengan menggunakan model MCNN pada Gambar 4 dan *Cascaded-MtL* dengan model arsitektur seperti pada Gambar 5 yang telah dimodifikasi dengan menggunakan python. *Training* dilakukan dengan menggunakan GPU NVIDIA GEFORCE GTX 1070 6GB. Proses *training* berlangsung selama kurang lebih 7 hari dengan menggunakan 2000 *epoch*, dan *learning rate* 0.000001.



Gambar 6: a Hasil estimasi density Part A; b Hasil estimasi density Part B

B. Pengujian

Pada bagian pengujian ini, akan digunakan gambar uji dari ShanghaiTech dataset yang memiliki spesifikasi 182 gambar uji coba untuk Part A dengan ukuran gambar yang beragam, untuk uji coba pada Part B digunakan 316 gambar uji coba dengan spesifikasi 768×1024 . Hasil dari proses merupakan *density map*. Disertai hasil evaluasi model yang berbentuk *Mean Absolute Error* (MAE) dan *Mean Square Error* (MSE). Pada pengujian ini juga digunakan 14 gambar yang didapatkan dari Internet, berupa gambar keramaian yang berada pada wilayah sekitar Kota Surabaya.

1) Pengujian Model Pertama

Pengujian pertama dilakukan dengan menggunakan weight yang didapatkan dari hasil *training* menggunakan MCNN. Berikut merupakan *density* dari kedua model uji coba:



Gambar 7: Hasil density map beserta gambar asli pada Part A

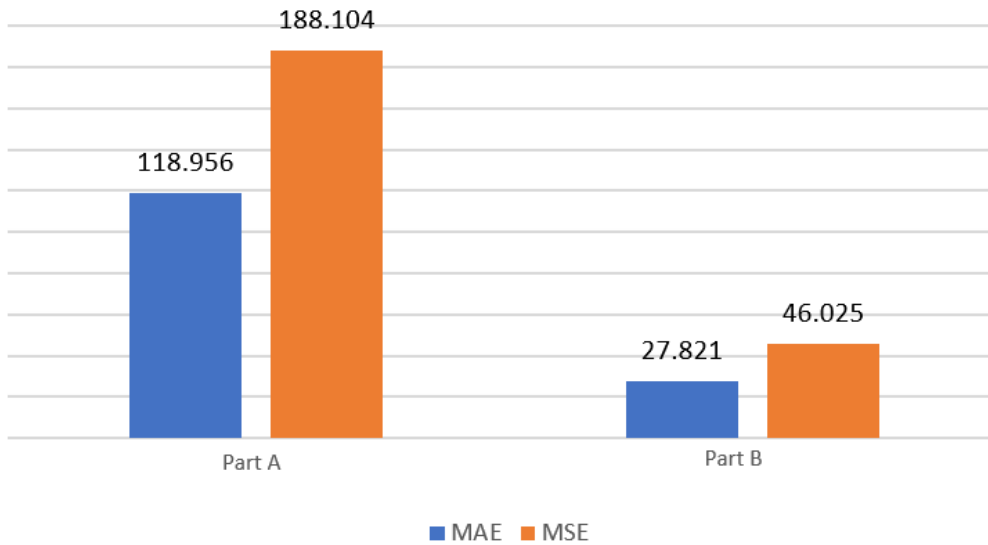


Gambar 8: Hasil density map beserta gambar asli pada Part B

Berikut merupakan tabel hasil dari *metrics validation* dengan menggunakan *Mean Absolut Error* (MAE) dan *Mean Square Error* (MSE).

Tabel 2: Hasil test pada Sanghaitech dataset

Part A				
Test	MAE	MSE	Test	MAE
1	118.67	187.97	6	118.74
2	120.42	188.97	7	120.00
3	118.60	187.82	8	118.72
4	117.87	187.58	9	118.15
5	120.35	188.99	10	118.40
Average		118.956	188.104	
Part B				
Test	MAE	MSE	Test	MAE
1	28.47	46.79	6	29.66
2	28.41	46.73	7	25.22
3	27.31	45.37	8	27.5
4	27.35	45.48	9	27.95
5	28.41	46.89	10	27.93
Average		27.821	46.025	



Gambar 9: Grafik nilai accuracy training model pertama

Dari grafik Gambar 9 menunjukkan bahwa nilai akurasi pada proses *training* pertama memiliki 2 hasil berbeda pada Part A dan Part B. Pada model ini didapatkan bahwa nilai akurasi sesuai dengan keadaan kepadatan yang terjadi pada gambar. Pada model pertama juga dilakukan pengujian dengan menggunakan 14 data gambar yang di peroleh dari internet dengan hasil seperti pada tabel berikut.

Tabel 3 menunjukkan hasil estimasi yang menggunakan 14 gambar dengan menyertakan *Mean Absolut Error* dari setiap gambar yang digunakan sebagai pengujian yang menggunakan hasil proses *training* dari model pertama. Dari data tersebut bisa kita lihat bahwa model ini berhasil melakukan estimasi terhadap beberapa gambar pengujian mendekati jumlah asli pada gambar tersebut yang berupa *ground truth*.

Tabel 3: Hasil Estimasi MCNN

Image	<i>ground truth</i>	Estimasi MCNN	MAE
taman bungkul	726	507.79	218.21
cfid 1	600	540.56	59.44
cfid 2	65	264.75	199.75
cfid 3	185	170.49	14.51
cfid 4	265	297.55	32.55
cfid 5	210	176.19	33.81
konser dewa19	106	74.21	31.79
grandcity 7	115	302.45	187.45
grandcity 9	87	626.14	539.14
grandcity 10	140	70.13	69.87
grandcity 11	78	238.11	160.11
kn 12	335	1041.15	706.15
kn 13	362	1115.41	753.41
kn 14	372	1544.62	1217.62

2) Pengujian Model Kedua

Pengujian model ke dua dilakukan dengan menggunakan metode yang sama dengan model pertama didapatkan dari hasil *training* menggunakan *Cascaded-MtL*. Berikut merupakan contoh hasil *density* dari model uji coba



Gambar 10: Hasil density map pada Cascaded-MtL beserta gambar asli

Pada model kedua juga dilakukan pengujian dengan menggunakan 14 data gambar yang di peroleh dari internet dengan hasil seperti pada tabel berikut.

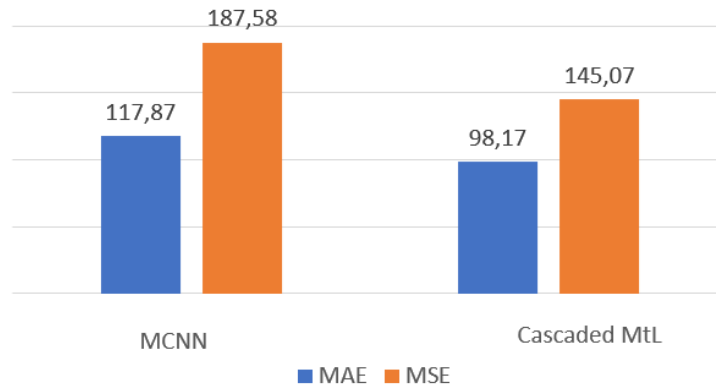
Tabel 4: Hasil Estimasi CMTL

Image	<i>ground truth</i>	Estimasi CMTL	MAE
taman bungkul	726	507.79	247.88
cfid 1	600	540.56	151.13
cfid 2	65	264.75	75.72
cfid 3	185	170.49	8.19
cfid 4	265	297.55	5.7
cfid 5	210	176.19	65.11
konser dewa19	106	74.21	18.16
grandcity 7	115	302.45	138.44
grandcity 9	87	626.14	326.09
grandcity 10	140	70.13	3.02
grandcity 11	78	238.11	109.39
kn 12	335	1041.15	523.94
kn 13	362	1115.41	517.47
kn 14	372	1544.62	926.4

Tabel 4 menunjukkan hasil Estimasi dari *Cascaded Multi-task Learning* yang juga menggunakan 14 gambar dengan menyertakan *Mean Absolut Error* dari setiap gambar yang digunakan sebagai pengujian, menggunakan hasil proses *training* dari model tersebut. Pada data tersebut bisa kita lihat bahwa model ini juga berhasil melakukan estimasi terhadap beberapa gambar pengujian mendekati jumlah asli pada gambar tersebut.

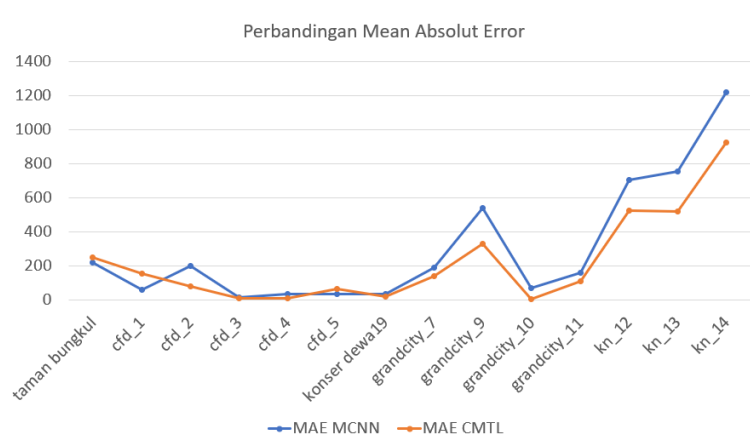
C. Analisa

Setelah dilakukan uji coba performa dengan menghadirkan MAE dan MSE pada ke dua model tersebut, didapatkan nilai rata - rata performa dari tiap model.



Gambar 11: Grafik nilai error pada training model pertama

Dari Gambar 11 menunjukkan bahwa nilai akurasi pada proses *training* dengan menggunakan Shanghaitech dataset pada model ke 2 memiliki keunggulan hasil dibandingkan dengan model pertama. Berikutnya dilakukan uji coba dengan menggunakan 14 gambar yang didapat dari Internet sesuai dengan Tabel 3 untuk estimasi dari MCNN dan Tabel 4 untuk estimasi dari CMTL.



Gambar 12: Grafik nilai perbandingan error pada kedua model

Pada Gambar 12 menunjukkan perbandingan bahwa model ke kedua lebih memiliki keunggulan pada estimasi gambar, yaitu lebih mendekati dari *ground truth* atau jumlah objek asli pada gambar yang di gunakan untuk melakukan pengujian.

V. KESIMPULAN

Berdasarkan hasil pengujian yang telah dilakukan, dapat ditarik beberapa kesimpulan sebagai berikut:

1. Berdasarkan hasil *training* dan merujuk pada Gambar 6, model pertama memperoleh nilai yang cukup baik pada Part B. Hal ini dikarenakan tingkat kepadatan objek pada gambar tersebut cenderung lebih rendah. Namun, perbandingan performa menunjukkan bahwa model kedua memiliki akurasi yang lebih baik. Keunggulan ini disebabkan oleh arsitektur *Cascaded-Mtl* yang memiliki fitur layer *high-level prior* untuk mengoptimalkan hasil estimasi.
2. Karakteristik latar tempat dan fitur objek pada dataset sangat memengaruhi hasil estimasi. Objek-objek dengan tekstur serupa kepala manusia, seperti dedaunan yang rimbun atau penggunaan aksesoris seperti topi, dapat menyebabkan kesalahan estimasi (*false positive*) selama proses pengujian.

Daftar Pustaka

- [1] Z. Liu, Y. Chen, and K. Xie, "Research on the impact of crowd flow on crowd risk in large gathering spots," in *2016 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII)*. IEEE, 2016.
- [2] H. Celik, A. Hanjalic, and E. A. Hendriks, "Towards a robust solution to people counting," in *2006 International Conference on Image Processing*. IEEE, 2006.
- [3] J. Xing et al., "Robust crowd counting using detection flow," in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011.
- [4] L. Wang, L. Wang, and F. Yang, "Counting crowd with fully convolutional networks," in *2017 2nd International Conference on Multimedia and Image Processing (ICMIP)*. IEEE, 2017.
- [5] 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, 2016.