ANOTASI OTOMATIS PADA CITRA PERSEPSI AUTONOMOUS CAR MENGGUNAKAN TRANSFER LABEL 3D-TO-2D

Areta Vahtsa Nur Kirana*[1], Muhammad Idham Ananta Timur[2] [2]
[1,2]Departemen Ilmu Komputer dan Elektronika, FMIPA UGM, Yogyakarta, Indonesia
e-mail: ***[1]areta.vahtsa2401@mail.ugm.ac.id**, [2]**idham@mail.ugm.ac.id**

Abstrak

Pemetaan semantik dan panoptik pada citra kendaraan otonom sangat bergantung pada ketersediaan anotasi manual yang memadai. Namun, proses pelabelan manual memerlukan waktu dan biaya tinggi, serta rentan terhadap inkonsistensi antar frame. Penelitian ini mengusulkan pendekatan otomatis berbasis Neural Radiance Fields(NeRF) untuk mentransfer label dari representasi 3D ke citra 2D. Dataset KITTI-360 digunakan dengan 768 citra yang dianotasi secara semi-otomatis menggunakan segmentor PSPNet dan Mask2Former. Model dilatih secara terpisah pada tiga sequence berbeda dan dievaluasi menggunakan metrik Intersection over Union (IoU), mean IoU (mIoU), dan Panoptic Quality (PQ). Hasil menunjukkan model mampu mencapai mIoU sebesar 0.79 (perspektif) dan 0.73 (fisheye), serta PQ sebesar 0.66 dan 0.60. Hasil kualitatif menunjukkan segmentasi yang konsisten pada kelas dominan seperti jalan, bangunan, dan langit, serta kemampuan membedakan instance objek dengan baik. Pendekatan ini terbukti mampu menghasilkan anotasi berkualitas tinggi tanpa ketergantungan pada pelabelan manual. Temuan ini penting untuk mempercepat pembangunan sistem persepsi visual kendaraan otonom dan memperluas cakupan data pelatihan secara efisien.

Kata kunci*—anotasi otomatis, neural radiance field, segmentasi panoptik, kitti-360, kendaraan otonom*

Abstract

Semantic and panoptic mapping in autonomous vehicle imagery heavily relies on high-quality manual annotations. However, manual labeling is time-consuming, costly, and prone to inconsistency across frames. This study proposes an automated labeling approach based on Neural Radiance Fields (NeRF) to transfer labels from 3D representations to 2D images using the PanopticNeRF-360 architecture. The KITTI-360 dataset was utilized, with 768 images annotated semi-automatically using PSPNet and Mask2Former segmentors. The model was trained on three separate sequences and evaluated using Intersection over Union (IoU), mean IoU (mIoU), and Panoptic Quality (PQ) metrics. Results show that the model achieved an mIoU of 0.79 (perspective) and 0.73 (fisheye), and a PQ of 0.66 and 0.60 respectively. Qualitative results indicate consistent segmentation on dominant classes such as roads, buildings, and skies, as well as accurate instance separation. This approach proves effective in generating high-quality annotations without reliance on manual labeling. The findings are significant for accelerating the development of autonomous vehicle perception systems and enabling scalable dataset generation.

Keywords—*automated annotation, neural radiance field, panoptic segmentation, kitti-360, autonomous vehicles*

1. PENDAHULUAN

Perkembangan teknologi autonomous driving menuntut sistem persepsi visual yang mampu mengenali objek, jalan, dan kondisi lingkungan dengan akurasi tinggi. Kualitas sistem ini sangat bergantung pada ketersediaan dataset yang lengkap dan akurat. Sayangnya, mayoritas proyek vision gagal karena keterbatasan data anotasi manual, baik dari sisi kualitas maupun kuantitas [1]. Pengumpulan label manual membutuhkan waktu, biaya, dan keahlian yang tinggi, sehingga menjadi kendala utama dalam pengembangan model.

Beberapa upaya telah dilakukan untuk mengatasi keterbatasan ini, seperti augmentasi data [2], penggunaan simulator seperti CARLA [3] dan LGSVL [4], serta metode pembelajaran *semi-supervised* [5]. Namun, pendekatan tersebut memiliki keterbatasan seperti distorsi citra akibat augmentasi [17], perbedaan domain simulasi dengan kenyataan, dan kebutuhan akan data label tetap menjadi tantangan. Alternatif menjanjikan yang muncul adalah otomatisasi anotasi citra menggunakan metode *machine learning*, termasuk *self-supervision* [18], kombinasi dengan *Large-Language Model* [20], dan *active learning* [21].

Beberapa penelitian terdahulu yang relevan antara lain TADAP [6], *Panoptic Neural Fields* [7], [16]. Terdapat metode lain seperti transfer label 3D-to-2D berbasis CRF [8], *weak-supervised object localization* [19], dan segmentasi dengan kamera RGB-D [15]. Penelitian-penelitian tersebut menjadi pijakan penting dalam mengembangkan pendekatan anotasi otomatis berbasis geometri 3D dan konsistensi spasial.

Penelitian ini bertujuan untuk mengembangkan metode anotasi otomatis pada citra persepsi kendaraan otonom menggunakan pendekatan transfer label dari representasi 3D ke 2D. Pendekatan ini mengintegrasikan data *bounding box* 3D kasar, *pseudo-label* 2D dari model pretrained, serta teknik Neural Radiance Field (NeRF) untuk merepresentasikan scene secara spasial. Evaluasi dilakukan menggunakan metrik *Intersection over Union* (IoU), *mean IoU* (mIoU), dan *Panoptic Quality* (PQ).

2. METODE PENELITIAN

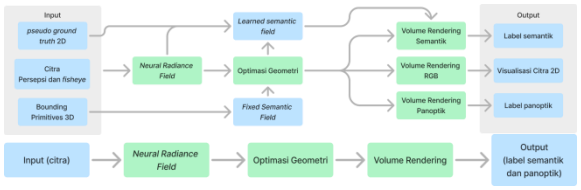
2.1 Akuisisi dan Persiapan Data

Tahapan awal dalam penelitian ini adalah akuisisi dan persiapan data, yang menjadi fondasi dari seluruh proses pengembangan model. Dataset yang digunakan adalah KITTI-360, sebuah dataset komprehensif yang menyediakan data citra dan sensor untuk lingkungan urban. Dataset ini memiliki keunggulan karena menyediakan citra dari kamera stereo (forward-facing) dan kamera fisheye dengan cakupan sudut 360°, memungkinkan simulasi kondisi sebenarnya di jalanan [10]. Pada gambar 1, dari dataset tersebut, dipilih sebanyak 768 citra, yang terdiri dari 384 citra perspektif dan 384 citra fisheye. Untuk proses pelatihan, digunakan pseudo ground truth yang dibangkitkan menggunakan model segmentasi semantik yaitu PSPNet [11] untuk citra perspektif, dan Mask2Former [12] untuk citra fisheye. Sementara itu, sebanyak 60 citra dengan anotasi manual digunakan secara khusus untuk tahap evaluasi akhir. Proses ini memastikan bahwa data latih dan data uji terpisah dan representatif terhadap variasi lingkungan.

Gambar 1. Citra yang diambil dari KITTI-360

2.2 Arsitektur Model

Model yang digunakan dalam penelitian ini mengadopsi arsitektur PanopticNeRF-360 [7], yang memanfaatkan keunggulan representasi NeRF [9] dalam menangkap struktur spasial dan semantik dari scene 3D, seperti yang terlihat pada Gambar 2. Model ini dirancang untuk menerima input berupa citra RGB, pseudo ground truth 2D, serta bounding primitives 3D yang bersifat kasar, seperti kubus atau elipsoid, yang mewakili objek dalam scene. Proses dimulai dari ray sampling, yaitu penembakan sinar dari kamera ke scene 3D berdasarkan parameter intrinsik dan ekstrinsik kamera. Titik-titik potong antara ray dan objek 3D dihitung melalui mesh-ray intersection. Informasi dari titik-titik ini kemudian dimasukkan ke dalam multilayer perceptron (MLP), yang menghasilkan output prediksi warna (RGB), kedalaman (depth), label semantik, serta ID instance. Hasil akhir berupa segmentasi panoptik memungkinkan sistem membedakan antara objek-objek unik (things) serta area luas homogen (stuff), seperti jalan atau langit.



Gambar 2. Gambar alur kerja model

2.3 Konfigurasi Training

Untuk melatih model agar dapat menghasilkan anotasi yang akurat, dilakukan konfigurasi pelatihan yang spesifik dan disesuaikan dengan karakteristik dataset. Pelatihan dilakukan secara terpisah pada tiga subset sequence, yang masing-masing mewakili kondisi jalanan yang berbeda, seperti daerah pepohonan, kawasan padat bangunan, dan area kombinasi. Pada gambar 3, setiap proses pelatihan menggunakan batch size sebesar 1 untuk memaksimalkan detail tiap frame, dengan jumlah epoch sebanyak 30, dan learning rate sebesar 5e-4. Parameter lainnya seperti jumlah ray yang disampling, ukuran chunk, dan kedalaman maksimum untuk ray tracing juga diatur sesuai dengan kapasitas perangkat keras dan kompleksitas scene. Pendekatan ini memungkinkan model belajar dengan efektif tanpa overfitting pada satu jenis lingkungan tertentu.

```yaml
train:
    batch_size: 1
    lr: 5e-4
    weight_decay: 0.
    epoch: 30
    scheduler:
        type: 'exponential'
        gamma: 0.5
        decay_epochs: 30
    num_workers: 3

test:
    dataset: KITTI360
    batch_size: 1

ep_iter: 1000
save_ep: 1
eval_ep: 1
save_latest_ep: 1
log_interval: 1
```

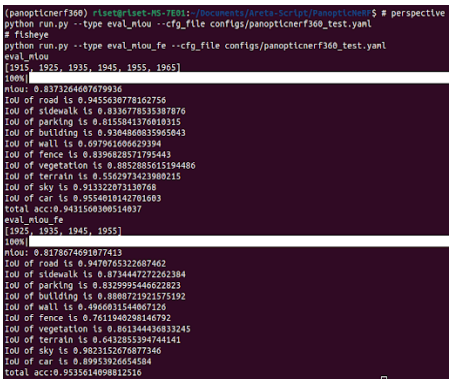
Gambar 3. Gambar alur kerja model

2.4 Penghitungan Loss

Selama pelatihan, model mengoptimalkan tiga jenis loss utama yang mewakili tugas-tugas penting dalam persepsi visual: rekonstruksi warna, estimasi kedalaman, dan segmentasi semantik. Pertama, RGB loss dihitung dengan fungsi Mean Squared Error (MSE) antara prediksi citra dan citra asli, untuk memastikan hasil render NeRF mendekati tampilan nyata. Kedua, depth loss menggunakan fungsi Huber Loss [14], yang memberikan penalti terhadap prediksi kedalaman yang meleset namun tetap robust terhadap outlier, dan diterapkan hanya pada piksel yang memiliki ground truth depth. Ketiga, semantic loss merupakan gabungan dari supervisi 2D dari pseudo ground truth, perbaikan label melalui fix loss, serta konsistensi semantik berdasarkan bounding primitives 3D. Ketiga loss ini dikombinasikan dengan bobot tertentu yang telah disesuaikan, dan dijumlahkan untuk membentuk total loss yang digunakan selama proses training. Strategi ini membantu model memahami hubungan antara informasi visual 2D dan struktur spasial 3D secara simultan.

2.5 Evaluasi

Untuk menilai performa model, dilakukan evaluasi terhadap citra yang telah dilabeli secara manual dari dataset KITTI-360, seperti yang terlihat pada gambar 4. Evaluasi ini dilakukan terhadap 60 citra (30 citra perspektif dan 30 citra fisheye) dengan membandingkan hasil prediksi model terhadap ground truth menggunakan metrik Intersection over Union (IoU), mean IoU (mIoU), dan Panoptic Quality (PQ) [13]. IoU digunakan untuk mengukur seberapa baik model mengenali batas objek, mIoU memberikan gambaran rata-rata akurasi segmentasi seluruh kelas, dan PQ menggabungkan kualitas segmentasi semantik dan panoptik. Evaluasi dilakukan secara terpisah untuk label semantik dan panoptik guna memahami kinerja model dalam mengenali jenis objek serta membedakan antar-instance dari objek yang sama.

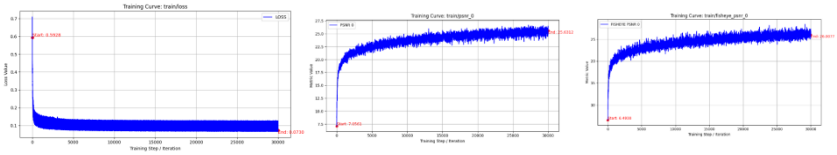


Gambar 4. Proses evaluasi model

3. HASIL DAN PEMBAHASAN

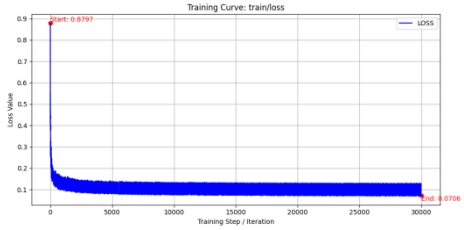
3.1 Hasil Training

Pada tahapan pelatihan, model diuji pada tiga subset citra berbeda untuk mengukur kemampuannya dalam menggeneralisasi scene yang bervariasi.



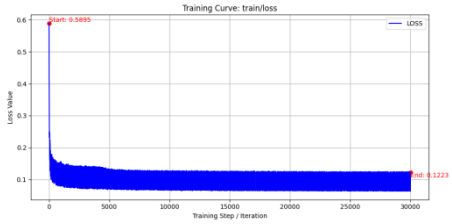
Gambar 5. a) *Train Loss* pada sequence 1728-1791, b) PSNR pada sequence 1728-1791

Pada gambar 5a, terdapat train loss dari sequence pertama (1728–1791), yang menggambarkan area dengan vegetasi cukup banyak di satu sisi dan deretan rumah di sisi lain, memperlihatkan penurunan nilai train loss dari 0.5928 menjadi 0.0730. Kurva penurunan loss yang mulus dan konsisten menandakan bahwa model mampu menyesuaikan diri terhadap data secara bertahap. Hal ini juga tercermin dalam nilai PSNR (Peak Signal-to-Noise Ratio) pada gambar 3.1b yang meningkat dari 7.05 dB menjadi lebih dari 25 dB, menunjukkan peningkatan kualitas visual yang signifikan.



Gambar 6. *Train Loss* pada sequence 1908–1971

Sementara itu, sequence kedua (1908–1971) pada gambar 6 memperlihatkan lingkungan yang lebih padat dan kompleks, dengan lebih banyak kendaraan terparkir serta variasi struktur bangunan di kedua sisi jalan. Train loss pada subset ini dimulai dari 0.8797 dan secara bertahap turun hingga 0.0706. Meskipun memiliki kompleksitas lebih tinggi, model tetap mampu mempertahankan tren penurunan loss yang stabil, membuktikan bahwa arsitektur PanopticNeRF-360 cukup fleksibel untuk menangani beragam jenis lingkungan visual.

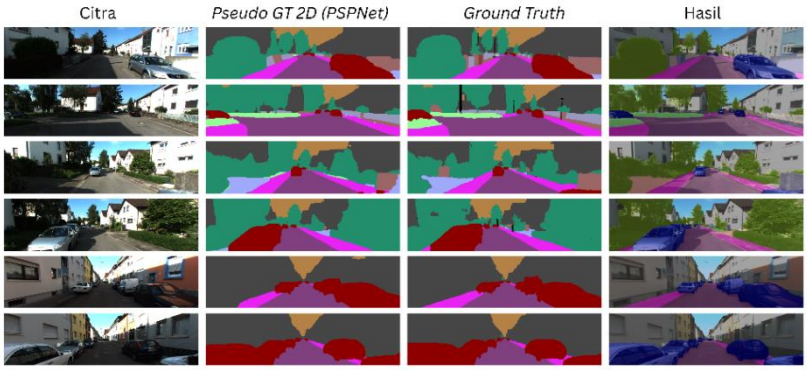


Gambar 7. *Train Loss* pada sequence 6398–6461

Gambar 7 memperlihatkan performa model pada sequence ketiga (6398–6461), citra menggambarkan kawasan urban dengan dominasi bangunan dan lebih sedikit vegetasi. Train loss awal sebesar 0.5895 menurun hingga mencapai 0.1223. Meskipun nilai akhir sedikit lebih tinggi dibanding dua sequence sebelumnya, hal ini masih dalam kategori baik, mengingat jenis objek yang lebih kompleks dan batas objek yang tidak terlalu kontras. Performa ini menandakan bahwa model memerlukan perhatian lebih pada segmentasi objek kecil atau saling berhimpitan.

3.2 Evaluasi Semantic Label

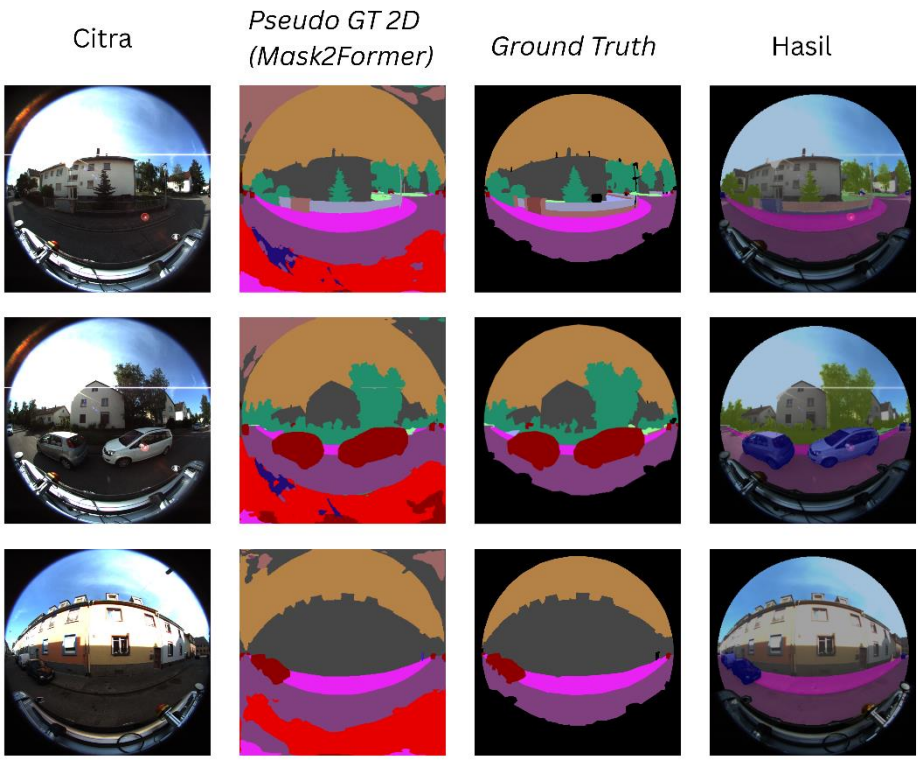
Visualisasi pada gambar 8 menunjukkan hasil prediksi model pada citra perspektif. Dari kolom Hasil, terlihat bahwa model berhasil mengidentifikasi area utama dengan cukup akurat, seperti jalan (road), langit (sky), bangunan (building), dan kendaraan (car). Pola warna pada prediksi sangat menyerupai ground truth manual (kolom ketiga), menandakan bahwa representasi semantik berhasil dipelajari dengan baik oleh model.



Gambar 8. Hasil Kualitatif Label Semantik pada Citra *Forward-facing*

Prediksi model pada kelas jalan konsisten mengikuti geometri jalan, bahkan pada bagian yang menyempit atau melengkung. Langit diidentifikasi secara utuh meskipun terpotong oleh dedaunan, dan bangunan dapat dikenali meski berada dalam kondisi pencahayaan berbeda. Model juga mampu mengatasi noise dari pseudo GT 2D (kolom kedua), di mana PSPNet menunjukkan kontur yang lebih kasar atau salah label, namun berhasil dikoreksi oleh prediksi akhir. Namun, masih terdapat sedikit kesalahan prediksi pada kelas seperti pagar (fence) dan box, yang kadang tercampur dengan bangunan atau tanah. Hal ini umum terjadi ketika objek memiliki ukuran kecil atau tepi yang tidak jelas. Secara keseluruhan, hasil semantik pada citra perspektif menunjukkan bahwa model dapat melakukan generalisasi visual dengan baik terhadap scene urban kompleks.

Pada gambar 9 yang menunjukkan visualisasi hasil model untuk citra fisheye, model juga menunjukkan kemampuan yang cukup baik dalam mengenali area luas seperti jalan, langit, dan bangunan. Bentuk prediksi yang lebih membulat mengikuti bentuk circular projection dari kamera menunjukkan bahwa model mampu mengadaptasi pada geometri distorsi khas fisheye.



Gambar 9. Hasil Kualitatif Label Semantik pada Citra *Fisheye*

Prediksi jalan dan langit tampak sangat konsisten pada seluruh baris. Ini penting karena fisheye kerap mengalami distorsi ekstrem di bagian tepi, namun model tetap dapat mempertahankan segmentasi yang utuh. Mobil-mobil yang berada di sisi kiri dan kanan citra juga berhasil dipisahkan dari latar, menunjukkan bahwa representasi 3D berhasil membantu membedakan elevasi dan kontur objek. Perbandingan dengan pseudo GT dari Mask2Former (kolom kedua) memperlihatkan bahwa model berhasil menghilangkan noise serta memperhalus batas objek, misalnya pada semak-semak dan kontur rumah. Namun, kesalahan kecil masih terlihat pada area bayangan atau objek tumpang tindih, di mana model kadang menyatukan dua kelas menjadi satu. Secara umum, performa model pada citra fisheye sangat kompeten meski terdapat tantangan dari distorsi lensa, dengan peningkatan kualitas visual dan semantik dibanding pseudo GT.

Evaluasi terhadap hasil segmentasi semantik dilakukan dengan membandingkan label prediksi model dengan label ground truth menggunakan metrik IoU dan mIoU, seperti yang terlihat pada tabel 1.

Tabel 1 Hasil Kuantitatif dari Label Semantik

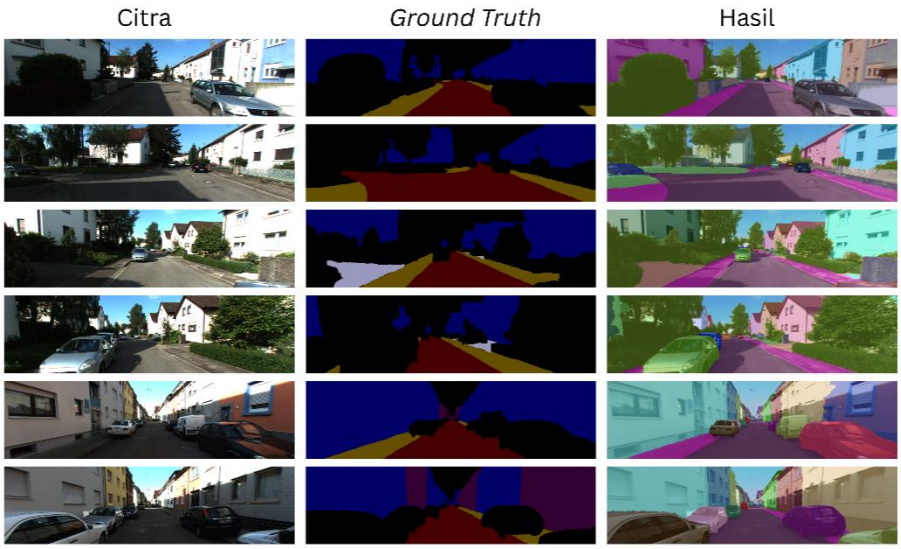
Forward-facing							
IoU						**mIoU**	**Acc**
Road	**Park**	**Sdwlk**	**Terr**	**Bldg**	**Vegt**		
0.96	0.79	0.85	0.65	0.94	0.88	0.79	0.86
Car	**Gate**	**Wall**	**Fence**	**Box**	**Sky**		

0.93	0.72	0.66	0.74	0.65	0.91		
Fisheye							
IoU						**mIoU**	**Acc**
Road	**Park**	**Sdwlk**	**Terr**	**Bldg**	**Vegt**		
0.96	0.63	0.88	0.56	0.91	0.6	0.73	0.77
Car	**Gate**	**Wall**	**Fence**	**Box**	**Sky**		
0.91	0.77	0.49	0.69	0.54	0.96		

Untuk citra perspektif, model mencapai mIoU sebesar 0.79, sedangkan untuk citra fisheye mencapai 0.73. Kelas dengan skor tertinggi di antaranya adalah road (0.96), building (0.94), car (0.93), dan sky (0.91), yang merupakan kelas dominan dalam scene urban. Sebaliknya, kelas seperti fence, box, dan gate memperoleh nilai IoU lebih rendah, yakni di bawah 0.70, menunjukkan tantangan dalam membedakan batas objek yang kecil atau terletak di tepi gambar. Citra fisheye cenderung memiliki distorsi pada bagian pinggir, yang menurunkan ketepatan prediksi pada area tersebut. Namun demikian, hasil ini tetap menunjukkan bahwa model cukup andal dalam memahami struktur utama scene dari berbagai sudut pandang.

3.3 Evaluasi Panoptic Label

Visualisasi pada gambar 10 menunjukkan hasil label panoptik pada kamera perspektif, memperlihatkan kemampuan model dalam mengenali dan membedakan objek berdasarkan kelas dan instance. Pada kolom “Hasil”, terlihat bahwa setiap mobil, bangunan, atau pagar yang unik berhasil diberi warna berbeda, menandakan pemisahan antar instance dalam satu kelas berhasil dilakukan.

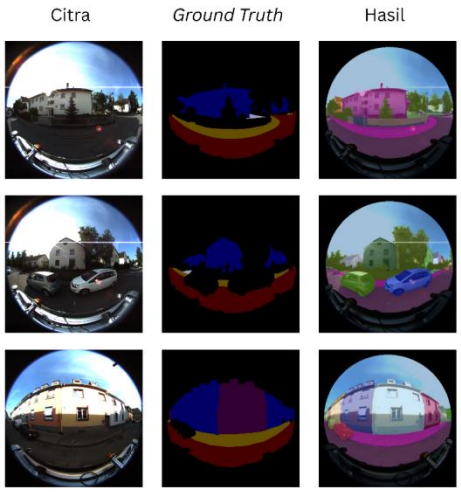


Gambar 10. Hasil Kualitatif Label Semantik pada Citra *Forward-facing*

Objek seperti mobil yang berada berdekatan di sepanjang sisi jalan tersegmentasi secara individual tanpa tumpang tindih. Hal ini menunjukkan bahwa proses instance segmentation

bekerja dengan cukup baik pada objek berukuran sedang hingga besar. Pada kelas stuff seperti jalan, langit, dan bangunan, hasil segmentasi juga terlihat utuh dan konsisten. Namun, terdapat beberapa tantangan yang masih muncul. Misalnya, pada objek kecil seperti pagar atau tanaman hias, model kadang gagal membedakan instance yang bersebelahan, atau malah menggabungkannya sebagai satu entitas. Di sisi lain, warna yang berbeda untuk setiap mobil menandakan keberhasilan model dalam membedakan mobil yang sejenis tapi berbeda identitas, yang sangat penting untuk konteks navigasi kendaraan otonom. Secara keseluruhan, hasil panoptik pada kamera perspektif menunjukkan bahwa pendekatan 3D-to-2D ini efektif dalam memahami struktur dan keberadaan objek secara detail dari satu arah pandang tetap.

Hasil panoptik untuk kamera fisheye pada gambar 11 menunjukkan kemampuan serupa dalam segmentasi instance, meskipun dengan tantangan tambahan berupa distorsi tepi akibat sudut pandang 360°. Dalam beberapa baris citra, terlihat bahwa beberapa kendaraan yang berada dekat lensa berhasil dipisahkan dengan baik dan diberi warna berbeda. Hal ini mencerminkan keberhasilan model dalam mengadaptasi input dari kamera non-konvensional.



Gambar 11. Hasil Kualitatif Label Semantik pada Citra *Forward-facing*

Area jalan dan langit tetap dikenali dengan sangat baik, dengan segmentasi stuff yang halus dan tidak terputus. Model juga mampu menghindari kesalahan pseudo GT, terutama pada bagian bawah citra yang sering kali mengalami clipping atau noise dari kendaraan atau bumper kamera. Kesalahan masih muncul pada bagian pinggir citra, di mana objek seperti pohon atau pagar terkadang tidak dipisahkan sempurna antar instance-nya, bahkan tidak dikenali sebagai objek terpisah. Walaupun demikian, secara umum instance segmentation tetap berhasil pada mayoritas objek things yang terlihat utuh dan cukup besar. Model tampaknya belajar dari konsistensi spasial 3D dan menghasilkan prediksi panoptik yang tidak hanya tajam secara semantik, tetapi juga tersegmentasi secara individual, bahkan dalam citra dengan sudut ekstrim.

Untuk mengevaluasi kemampuan model dalam menghasilkan label panoptik, digunakan metrik Panoptic Quality (PQ), yang menggabungkan kualitas segmentasi semantik dan identifikasi instance objek, seperti yang terlihat pada tabel 2. Nilai PQ keseluruhan untuk citra perspektif mencapai 0.66, sementara untuk citra fisheye sebesar 0.60. Secara umum, PQ-stuff (untuk objek homogen seperti jalan dan langit) memiliki nilai yang lebih tinggi dibandingkan PQ-things (untuk objek individual seperti mobil, pagar, dan box). Hasil ini menunjukkan bahwa model memiliki kecenderungan lebih kuat dalam mengenali area luas dengan label tunggal dibanding membedakan objek-objek kecil yang berdekatan.

Tabel 2 Hasil Kuantitatif dari Label Panoptik

Forward-facing								
PQ						**PQ All**	**PQ things**	**PQ stuff**
Road	**Park**	**Sdwlk**	**Terr**	**Bldg**	**Vegt**			
0.96	0.58	0.83	0.5	0.62	0.87			
Car	**Gate**	**Wall**	**Fence**	**Box**	**Sky**			
0.71	0.6	0.57	0.56	0.27	0.9	0.66	0.71	0.77
Fisheye								
PQ						**PQ All**	**PQ things**	**PQ stuff**
Road	**Park**	**Sdwlk**	**Terr**	**Bldg**	**Vegt**			
0.93	0.46	0.86	0.25	0.71	0.78			
Car	**Gate**	**Wall**	**Fence**	**Box**	**Sky**			
0.49	0.54	0.27	0.41	0.57	0.94	0.6	0.56	0.71

Meskipun hasil evaluasi menunjukkan performa yang menjanjikan, masih terdapat beberapa jenis kesalahan yang perlu diperhatikan. Salah satu kesalahan yang paling umum adalah over-segmentation pada objek kecil seperti box atau pagar, terutama jika objek tersebut berada di pinggir citra fisheye yang mengalami distorsi optik. Selain itu, model juga cenderung menggabungkan dua instance berbeda menjadi satu (under-segmentation), khususnya jika objek-objek tersebut saling berdekatan atau memiliki warna serta tekstur serupa.

4. KESIMPULAN

Kesimpulan yang dapat ditarik dari penelitian ini adalah model dapat melakukan anotasi pada citra persepsi berupa forward-facing dan fisheye dengan baik, dengan melihat nilai mean intersection unit dan panoptic quality yang di atas 0.5. Selain itu, model dapat menangkap detail objek dari citra yang underexposed (minim cahaya) maupun citra yang overexposed (terlalu banyak cahaya) dengan baik.

DAFTAR PUSTAKA

[1] Wakefield Research, "Synthetic Data: key to production – Ready AI in 2022," Datagen, 2021. Available: https://wakefieldresearch.com/

[2] N. Ahn, B. Kang, and K.-A. Sohn, "Image Distortion Detection using Convolutional Neural Network," *arXiv (Cornell University)*, Jan. 2018, doi: 10.48550/arxiv.1805.10881. Available: https://arxiv.org/abs/1805.10881

[3] "Tool for automatic labeling of objects in images obtained from CARLA Autonomous Driving Simulator," *IEEEE*, Jan. 2017, doi: 10.1109/zinc58345.2023.10174056. Available: https://doi.org/10.1109/zinc58345.2023.10174056

[4] G. Rong *et al.*, "LGSVL Simulator: A high fidelity simulator for autonomous driving," *arXiv (Cornell University)*, Jan. 2020, doi: 10.48550/arxiv.2005.03778. Available: https://arxiv.org/abs/2005.03778

[5] D. Bergmann, "Semi-Supervised Learning," *IBM*, Sep. 02, 2024. Available: https://www.ibm.com/topics/semi-supervised-learning. [Accessed: Oct. 02, 2024]

[6] E. Alamikkotervo, R. Ojala, A. Seppänen, and K. Tammi, "TADAP: Trajectory-Aided Drivable area Auto-labeling with Pre-trained self-supervised features in winter driving

conditions," *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2312.12954. Available: https://arxiv.org/abs/2312.12954

[7] X. Fu *et al.*, "PanopticNERF-360: Panoramic 3D-to-2D label transfer in urban Scenes," *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2309.10815. Available: https://arxiv.org/abs/2309.10815

[8] J. Lee, S. W. Lee, H. Hong, and J. W. Jeon, "Automatic segmentation labeling for autonomous driving datasets," *2022 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, Oct. 2023, doi: 10.1109/icce-asia59966.2023.10326373. Available: https://doi.org/10.1109/icce-asia59966.2023.10326373

[9] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NERF: Representing scenes as neural radiance fields for view synthesis," *arXiv (Cornell University)*, Jan. 2020, doi: 10.48550/arxiv.2003.08934. Available: https://arxiv.org/abs/2003.08934

[10] Y. Liao, J. Xie, and A. Geiger, "KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D," *arXiv (Cornell University)*, Jan. 2021, doi: 10.48550/arxiv.2109.13410. Available: https://arxiv.org/abs/2109.13410

[11] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing network," *arXiv (Cornell University)*, Jan. 2016, doi: 10.48550/arxiv.1612.01105. Available: https://arxiv.org/abs/1612.01105

[12] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," *arXiv (Cornell University)*, Jan. 2021, doi: 10.48550/arxiv.2112.01527. Available: https://arxiv.org/abs/2112.01527

[13] O. Hahn, C. Reich, N. Araslanov, D. Cremers, C. Rupprecht, and S. Roth, "Scene-Centric Unsupervised panoptic segmentation," *arXiv.org*, Apr. 02, 2025. Available: https://arxiv.org/abs/2504.01955

[14] K. Gokcesu and H. Gokcesu, "Generalized Huber Loss for Robust Learning and its Efficient Minimization for a Robust Statistics," *arXiv.org*, Aug. 28, 2021. Available: http://arxiv.org/abs/2108.12627v1

[15] A. Helnawan, M. Attamimi, and A. N. Irfansyah, "Sistem Segmentasi Jalan dan Objek untuk Kendaraan Otonom Menggunakan Kamera RGB-D," *Jurnal Teknik ITS*, vol. 12, no. 1, May 2023, doi: 10.12962/j23373539.v12i1.110848. Available: https://doi.org/10.12962/j23373539.v12i1.110848

[16] A. Kundu *et al.*, "Panoptic Neural Fields: a semantic Object-Aware neural scene representation," *arXiv (Cornell University)*, Jan. 2022, doi: 10.48550/arxiv.2205.04334. Available: https://arxiv.org/abs/2205.04334

[17] M. Ofori-Oduro and M. Amer, "Defending Object Detection Models against Image Distortions," *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3842–3851, Jan. 2024, doi: 10.1109/wacv57701.2024.00381. Available: https://doi.org/10.1109/wacv57701.2024.00381

[18] J. Seo, S. Sim, and I. Shim, "Learning Off-Road terrain traversability with Self-Supervisions only," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4617–4624, Jun. 2023, doi: 10.1109/lra.2023.3284356. Available: https://doi.org/10.1109/lra.2023.3284356

[19] X. Xie *et al.*, "Weakly supervised object localization with soft guidance and channel erasing for auto labelling in autonomous driving systems," *ISA Transactions*, vol. 132, pp. 39–51, Aug. 2022, doi: 10.1016/j.isatra.2022.08.003. Available: https://doi.org/10.1016/j.isatra.2022.08.003

[20] X. Zhang, Z. Zhou, D. Chen, and Y. E. Wang, "AutoDistill: an End-to-End Framework to Explore and Distill Hardware-Efficient Language Models," *arXiv (Cornell University)*, Jan. 2022, doi: 10.48550/arxiv.2201.08539. Available: https://arxiv.org/abs/2201.08539

[21] Y. Zhou, S. Zhang, C. Meng, J. Mei, J. Li, and M. Wang, "A compact 3D lane auto labeling method for autonomous driving," *2021 IEEE International Conference on*

Unmanned Systems (ICUS), pp. 1177–1183, Oct. 2023, doi: 10.1109/icus58632.2023.10318323. Available: https://doi.org/10.1109/icus58632.2023.10318323