Rancangan EDA

Video Dashcam | Dataset BDD100K

Gambaran Fungsi BLOK

Mengapa dibagi menjadi beberapa blok?

BLOK A	**Data Contract & Joinability** Memastikan dataset bisa dipakai: video terbaca, label terbaca, dan label bisa di-join ke video secara valid.
BLOK B	**Label Health & Data Quality (MOT-ready)** Audit kualitas anotasi: bbox valid, tidak duplikat fatal, tidak ada label rusak yang bikin training “bohong”.
BLOK C	**Distribution & Difficulty Profiling (Video-centric)** Memahami karakter data yang menentukan strategi model: imbalance class, objek kecil, frame ramai, track pendek, dll.
BLOK D	**Visual Sanity & Edge Cases** Validasi dengan mata: bbox bener nempel objek? koordinat kebalik? kondisi gelap/blur dominan?

BLOK A — Data Contract & Joinability

(Video-Level)

Fungsi BLOK A

Blok ini bertugas memastikan dataset benar-benar “layak dipakai” sebelum bicara kualitas, distribusi, atau modeling. Intinya menjawab pertanyaan fundamental ML Engineer:

- Apakah video bisa dibaca?
- Apakah label valid?
- Apakah label benar-benar bisa di-join ke video tanpa ambiguity?

Kalau BLOK A gagal → EDA berikutnya tidak sah secara teknis.

STEP A1 — Dataset Inventory & Folder Integrity

Tujuan

Memastikan seluruh aset dataset (video + label) tersedia, konsisten, dan tidak rusak.

Insight yang ingin dicari

- Berapa jumlah video train yang tersedia?
- Format video apa saja yang digunakan?
- Apakah ada file video abnormal (0 byte, sangat kecil, dll)?
- Apakah file label (csv & parquet) dapat dibaca normal?

Yang dieksplorasi

- Struktur folder bdd100k/videos/train
- Nama file & ekstensi video
- Ukuran file video
- Basic load mot_labels.csv dan mot_labels.parquet

Fungsi / tools

- os, pathlib, glob
- pandas.read_csv()
- pandas.read_parquet()

Output

- Tabel inventory video (filename, ext, size_MB)
- Ringkasan: total video & total ukuran data
- Status akses label (OK / error)

STEP A2 — Label Schema Understanding (MOT Contract)

□ Tujuan

Memahami kontrak label secara eksplisit agar tidak salah interpretasi saat preprocessing & training.

□ Insight yang ingin dicari

- Kolom apa saja yang tersedia dalam label?
- Kolom mana yang wajib (video_id, frame_id, bbox, class, track_id)?
- Format bbox yang digunakan (xywh / xyxy)?
- Apakah koordinat pixel atau normalized?
- Tipe data tiap kolom (int, float, string)?

□ Yang dieksplorasi

- df.head(), df.info()
- Range nilai frame index
- Nilai unik class / category
- Keberadaan identifier video / sequence

□ Fungsi / tools

- DataFrame.head(), info(), describe()
- nunique(), unique()

□ Output

- Schema card (kolom, tipe data, deskripsi interpretatif)
- Dugaan key utama label (candidate primary key)
- Keputusan format bbox & frame indexing

STEP A3 — Video ↔ Label Joinability Check

□ Tujuan

Memastikan label benar-benar bisa dipetakan ke file video secara deterministik.

□ Insight yang ingin dicari

- Apakah setiap label mengacu ke video yang benar-benar ada?
- Apakah ada video tanpa label?
- Apakah ada label “orphan” (video_id tidak ditemukan)?

□ Yang dieksplorasi

- Mapping video_id / sequence_name ↔ nama file video
- Coverage video berlabel vs tidak berlabel
- Konsistensi nama & identifier

□ Fungsi / tools

- Set operations (set())
- merge, isin()
- value_counts()

Output

- Coverage report (% video berlabel / tanpa label)
- Daftar mismatch (jika ada)
- Keputusan final source label (csv atau parquet)

STEP A4 — Early Assumption Validation (Fail Fast)

Tujuan

Menghindari asumsi salah sejak awal yang bisa merusak seluruh pipeline.

Insight yang ingin dicari

- Apakah satu video = satu sequence?
- Apakah frame index konsisten dan berurutan?
- Apakah satu track_id unik hanya dalam satu video?

Yang dieksplorasi

- Track_id muncul di berapa video
- Frame continuity per track
- Duplikasi key dasar (video, frame, track)

Fungsi / tools

- groupby()
- duplicated()
- Simple sanity checks berbasis rule

Output

- Daftar asumsi yang VALID
- Daftar asumsi yang INVALID / perlu penanganan khusus
- Catatan batasan dataset untuk step berikutnya

BLOK B — Label Health & Data Quality

(MOT-Ready)

Fungsi BLOK B

Blok ini bertugas mengaudit kualitas anotasi, bukan untuk “membersihkan data” dulu, tapi untuk:

- Apakah label cukup sehat untuk training detection / tracking?
- Masalah apa yang berisiko merusak learning signal?
- Apakah problem ini minor (bisa diabaikan) atau fatal (harus ditangani)?

Output BLOK B akan menjadi dasar keputusan cleaning, filtering, dan preprocessing.

STEP B1 — Missingness & Basic Validity Check

Tujuan

Mendeteksi masalah label paling dasar yang tidak boleh lolos ke tahap modeling.

Insight yang ingin dicari

- Apakah ada missing value pada kolom penting?
- Apakah ada bbox tidak valid (negatif, nol, atau absurd)?
- Apakah ada class / category tidak dikenal?

Yang dieksplorasi

- Missing value per kolom
- Nilai bbox: width / height ≤ 0, x / y negatif
- Class id / label di luar daftar resmi

Fungsi / tools

- isna().mean()
- Boolean masking rule-based
- value_counts()

Output

- Tabel ringkas “label issue summary”
- Contoh baris bermasalah (sample)
- Status masalah: minor / major / fatal

STEP B2 — Duplicate Annotation & Key Consistency

Tujuan

Mendeteksi duplikasi dan konflik kunci yang bisa membuat model “belajar dua kali dari objek yang sama”.

□ Insight yang ingin dicari

- Apakah ada anotasi duplikat?
- Apakah kombinasi (video, frame, track_id) unik?
- Apakah satu track memiliki lebih dari satu bbox di frame yang sama?

□ Yang dieksplorasi

- Duplicate row secara literal
- Duplicate pada key logis
- Konflik track_id dalam frame

□ Fungsi / tools

- duplicated()
- groupby().size()

□ Output

- Statistik duplikasi
- Daftar conflict key (jika ada)
- Rekomendasi treatment (drop / merge / ignore)

STEP B3 — Bounding Box Geometry Sanity

□ Tujuan

Memastikan geometri bbox masuk akal secara fisik pada frame video.

□ Insight yang ingin dicari

- Apakah bbox keluar dari batas frame?
- Apakah bbox terlalu kecil / terlalu besar secara ekstrem?
- Apakah aspect ratio tidak realistis?

□ Yang dieksplorasi

- Bbox melebihi resolusi frame
- Distribusi area bbox
- Distribusi aspect ratio (w/h)

□ Fungsi / tools

- Rule check berbasis width/height frame
- describe(), quantile
- Histogram area & aspect ratio

□ Output

- Statistik bbox invalid / ekstrem
- Threshold rekomendasi (clip / drop)
- Indikasi potensi masalah small object

STEP B4 — Track Continuity & Temporal Consistency

Tujuan

Mengevaluasi apakah anotasi tracking stabil secara temporal atau terlalu “patah-patah”.

Insight yang ingin dicari

- Berapa panjang rata-rata track?
- Banyak track sangat pendek (1–2 frame)?
- Apakah frame index track meloncat ekstrem?

Yang dieksplorasi

- Track length per (video, track_id)
- Distribusi panjang track
- Gap frame dalam track

Fungsi / tools

- groupby().frame.nunique()
- diff() pada frame index

Output

- Distribusi panjang track
- Daftar track ekstrem (terlalu pendek / terlalu panjang)
- Indikasi tingkat kesulitan tracking

STEP B5 — Label Noise & Ambiguity (Early Signal)

Tujuan

Mengidentifikasi indikasi label noise yang tidak terlihat lewat statistik sederhana.

Insight yang ingin dicari

- Apakah ada bbox tumpang tindih ekstrem dengan class berbeda?
- Apakah satu track berganti class di tengah jalan?
- Apakah class tertentu sering ambigu?

Yang dieksplorasi

- Track dengan multiple class
- Overlap bbox (IoU tinggi) antar class
- Distribusi perubahan class per track

Fungsi / tools

- groupby(track_id).nunique(class)
- Perhitungan IoU sederhana
- Conditional statistics

Output

- Daftar indikasi label noise
- Class yang paling ambigu
- Catatan risiko terhadap training

BLOK C — Distribution & Difficulty Profiling

(Video-Centric)

Fungsi BLOK C

Blok ini bertugas memetakan “tingkat kesulitan” dataset dari sudut pandang model. BLOK C menjawab pertanyaan krusial ML Engineer:

- Apakah data seimbang atau long-tail?
- Apakah objek dominan kecil?
- Apakah frame cenderung ramai atau sepi?
- Apakah tracking sulit karena track pendek & terputus?

Output BLOK C akan langsung memengaruhi: pemilihan model, input resolution, augmentasi, strategi sampling & balancing.

STEP C1 — Class Distribution (Object & Track Level)

Tujuan

Memahami ketidakseimbangan kelas pada level deteksi dan tracking.

Insight yang ingin dicari

- Kelas apa yang dominan?
- Kelas apa yang langka (long-tail)?
- Apakah distribusi bbox ≠ distribusi track?

Yang dieksplorasi

- Jumlah bbox per class
- Jumlah unique track per class
- Perbandingan object-count vs track-count

Fungsi / tools

- value_counts()
- groupby(class).nunique(track_id)
- Bar plot (EDA-only)

Output

- Tabel distribusi class (bbox & track)
- Identifikasi class minoritas
- Rekomendasi awal balancing / reweighting

STEP C2 — Object Density per Frame (Crowdedness)

□ Tujuan

Mengukur tingkat “keramaian” frame yang berdampak langsung pada NMS & inference.

□ Insight yang ingin dicari

- Frame rata-rata berisi berapa objek?
- Seberapa sering frame sangat padat?
- Apakah ada ekor panjang (extreme crowded scenes)?

□ Yang dieksplorasi

- Jumlah bbox per (video, frame)
- Statistik mean, p90, p99
- Distribusi crowded vs sparse frame

□ Fungsi / tools

- groupby(['video','frame']).size()
- Histogram / ECDF

□ Output

- Distribusi objects-per-frame
- Threshold “crowded frame”
- Rekomendasi max_det, NMS strategy

STEP C3 — Bounding Box Size Profiling

Menilai dominasi objek kecil yang sering menjadi bottleneck mAP.

STEP C4 — Aspect Ratio & Shape Bias

Mendeteksi bias bentuk objek yang berpengaruh ke anchor-based / anchor-free model.

STEP C5 — Temporal Difficulty (Track Length & Fragmentation)

Mengukur tingkat kesulitan tracking dari sisi durasi & fragmentasi objek.

BLOK D — Visual Sanity & Edge Case Diagnostics

Fungsi BLOK D

Blok ini bertugas memvalidasi hasil EDA statistik dengan pengamatan visual langsung. BLOK D menjawab pertanyaan krusial:

- Apakah bbox benar-benar “nempel” objek?
- Apakah ada pola error label yang berulang?
- Apakah kondisi sulit (gelap, blur, crowded) dominan?

Banyak masalah tidak terlihat dari angka, tapi langsung jelas saat divisualisasikan.

STEP D1 — Label Overlay Sanity Check (Qualitative)

Memastikan koordinat bbox, class, dan skala benar secara visual.

STEP D2 — Crowded Scene & Occlusion Visual Audit

Memahami dampak occlusion & crowding terhadap kualitas anotasi.

STEP D3 — Lighting Condition (Day / Night Proxy)

Mendeteksi dominasi kondisi gelap / terang tanpa label cuaca eksplisit.

STEP D4 — Motion Blur & Sharpness Proxy

Mengukur pengaruh blur (gerak cepat, kamera) terhadap kualitas label.

STEP D5 — Extreme / Failure Case Sampling

Mengidentifikasi kasus ekstrem yang kemungkinan besar jadi error utama model.

—— End of Document ——