Introducere și Motivație

Context și impact social:

- Interacțiunea om-mașină devine din ce în ce mai frecventă prin intermediul telefoanelor, asistenților virtuali și roboților sociali. Capacitatea sistemelor de a înțelege stările emoționale îmbunătățește experiența utilizatorului și permite răspunsuri adaptive, empatice.
- În psihologie și medicină, detecția automată a expresiilor faciale poate sprijini diagnosticul și monitorizarea stărilor precum depresia sau anxietatea, precum și studiile comportamentale.
- În domenii precum marketingul digital și realitatea augmentată, recunoașterea emoțiilor contribuie la personalizarea conținutului și la crearea de experiențe emoțional relevante.

De ce problema detecției expresiilor faciale:

- Expresiile faciale sunt unul dintre cele mai naturale semnale non-verbale ale emoțiilor umane și oferă informații esențiale despre intenții și stări interioare.
- Detecția facială în medii „în sălbăticie" (imagini necontrolate) reprezintă o provocare majoră datorită variațiilor de iluminare, poziției feței, rezoluției și zgomotului.
- Soluțiile existente au performanțe bune în condiții controlate, dar generează rezultate suboptime pe date din lumea reală.

Alegerea dataset-ului AffectNet:

- **Scalabilitate:** Peste 420.000 de imagini etichetate manual, mult peste alte surse concurente, oferind diversitate de fețe, vârste, etnii și condiții de captură.
- **Acoperire emoțională:** Include 8 clase distincte de expresii (neutru, fericit, supărat, trist, teamă, surpriză, dezgust, dispreț), plus intensități continue de valență și arousal, permițând extinderi viitoare spre regresie emoțională.
- **Relevanță practică:** Date colectate din rețele sociale și web, reprezentative pentru scenarii de utilizare din aplicații reale.
- **Format compatibil YOLO:** Varianta procesată pe Kaggle, redimensionată la 96×96 pixeli, permite antrenarea rapidă și eficientă pe hardware obișnuit.

Descrierea structurii dataset-ului YOLO

- **Structura directoarelor**:
    - train/:
        - images/: imagini .png 96×96 px pentru antrenare
        - labels/: fișiere .txt cu coordonatele normalizate (x_center, y_center, width, height) și ID-ul clasei pentru fiecare imagine
    - validation/:

        - images/: imagini .png 96×96 px pentru validare
        - labels/: fișiere .txt cu etichete corespondente
    - test/:
        - images/: imagini .png 96×96 px pentru testare finală
        - labels/: fișiere .txt cu etichete pentru evaluarea performanțelor
- data.yaml: definește căile către seturile train, validation și test, precum și maparea celor 8 clase de expresii.
- **Eșantionare echilibrată**: Pentru a evita dezechilibrele de clasă și a asigura o distribuție uniformă, setul de date a fost ulterior eșantionat în subseturi cu același număr de exemple per clasă (800 pentru antrenare și 200 pentru validare). Acest proces folosește o funcție dedicată care garantează acoperirea egală a tuturor celor 8 clase și amestecarea aleatorie, menținând astfel diversitatea și stabilitatea metricilor de performanță în timpul antrenării și validării.

Antrenarea Modelului

Antrenarea cu YOLOv8 (în exemplul de mai sus) a constat în mai mulți pași și ajustări pentru a obține un model robust, capabil să detecteze expresiile faciale din subseturile echilibrate:

1. **Încărcarea modelului de bază**

model1 = YOLO('yolov8m.pt')

Am pornit de la versiunea „medium” pre-antrenată pe COCO, care oferă un bun compromis între complexitate și performanță.

2. **Parametri principali de antrenare**
    - data='data/YOLO_subset/data.yaml'
      Indică fișierul YAML cu căile către folderele train, val și test, numărul de clase (8) și numele fiecăreia.
    - epochs=200
      Am ales 200 de epoci pentru a ne asigura că modelul convergă complet, în special atunci când folosim augmentări puternice.
    - batch=8
      Un batch mic, potrivit pentru GPU-uri cu VRAM limitat, asigură totuși stabilitate a gradientului datorită regularizării implicite a gradient noise.
    - imgsz=96
      Dimensiunea imaginilor menține detaliul facial necesar clasificării expresiilor.
    - patience=50
      Early stopping după 50 de epoci fără îmbunătățire a metricii de validare, pentru a evita supraînvățarea și a economisi timp de calcul.
3. **Augmentări puternice pentru generalizare**
    - augment=True activează pipeline-ul complet de augmentări YOLOv8.

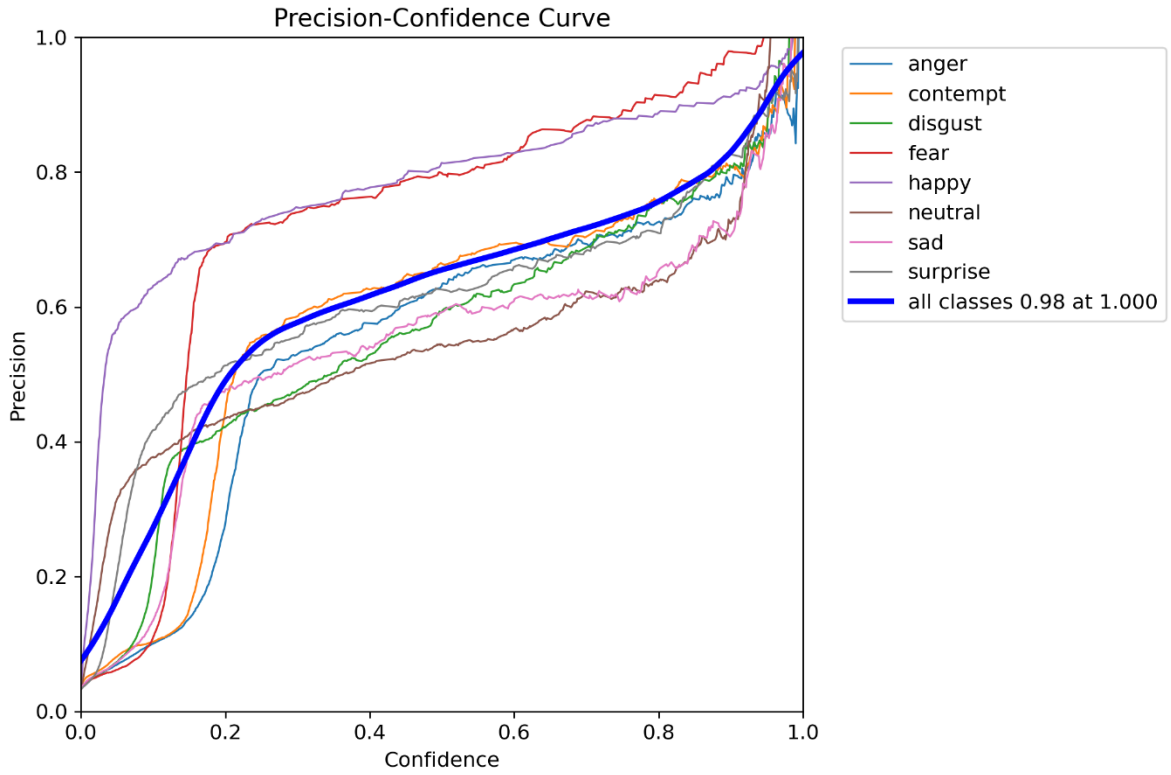
- mosaic=0.5 combină patru imagini într-una, cu probabilitate 50%, pentru a expune rețeaua la contexte variate și scale diferite.
- mixup=0.2 suprapune două imagini cu procent de 20%, ajutând la regularizare și la suavizarea hotarelor dintre clase.
- Ajustări **HSV** pentru schimbarea culorii:
  - hsv_h=0.015 (±1.5% pe nuanță)
  - hsv_s=0.7 (±70% pe saturație)
  - hsv_v=0.4 (±40% pe luminozitate)
    Aceste variații ajută modelul să fie robust la condiții diferite de iluminare și balans de alb.
- Flip-uri aleatorii:
  - flipud=0.2 (20% șansă de flip vertical)
  - fliplr=0.5 (50% șansă de flip orizontal)
    Aceste operațiuni cresc variabilitatea pozițiilor feței.

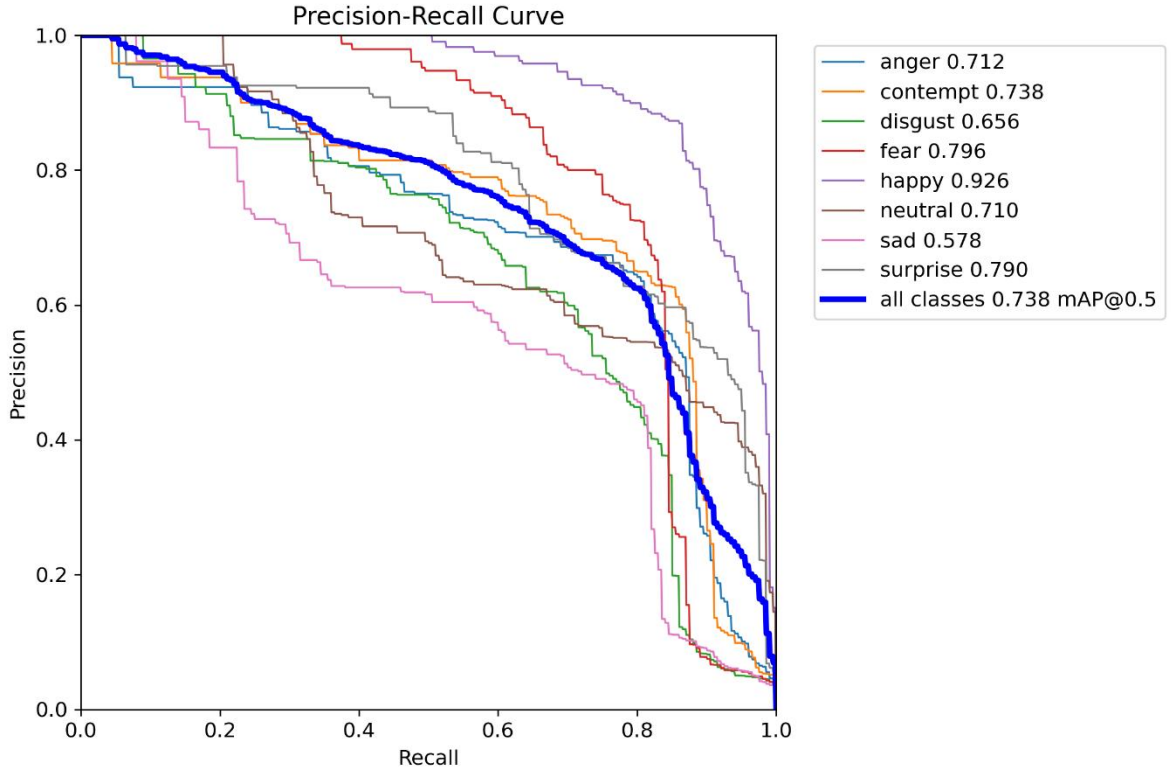
Ce rezultă din această abordare?

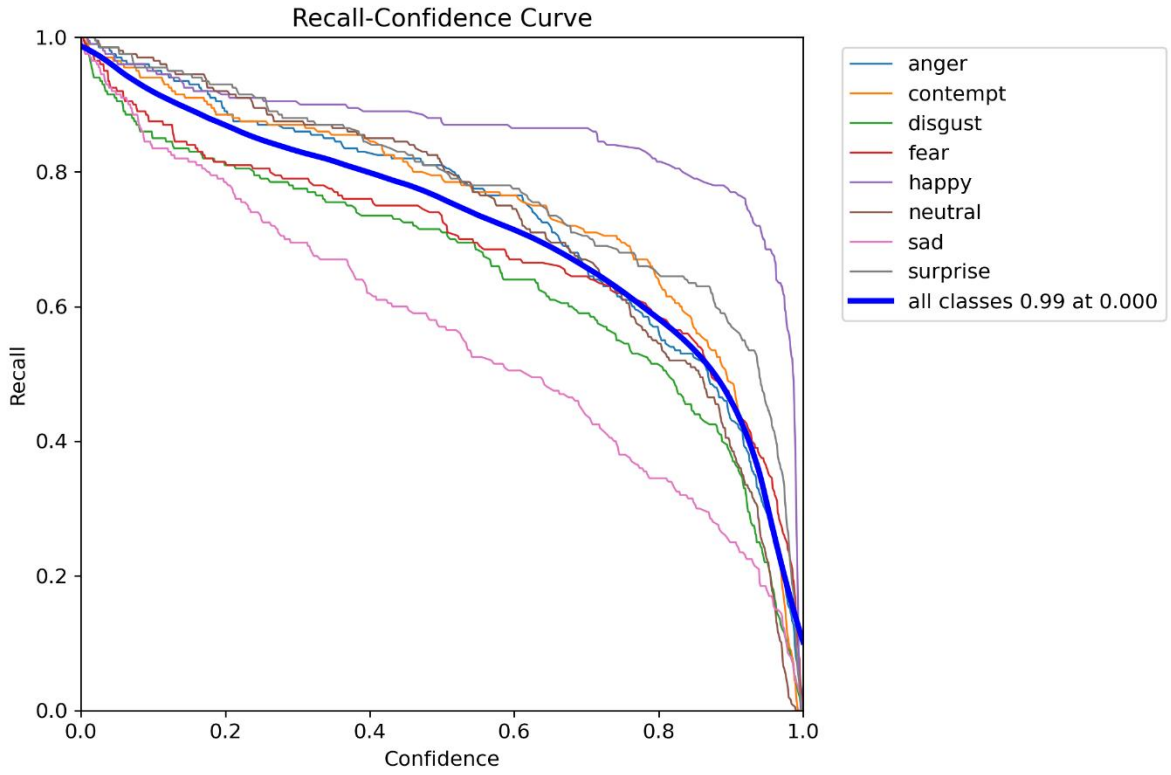
- Un model care învață atât din structura feței (datorită imagsz=96), cât și din variații puternice de context și culoare (augmentările MOSAIC, MIXUP, HSV).
- Early stopping previne supraspecializarea pe subset și ajută la alegerea greutăților optime pentru inferență.
- Un batch mic și epoci suficient de multe asigură convergență chiar și cu VRAM limitat.

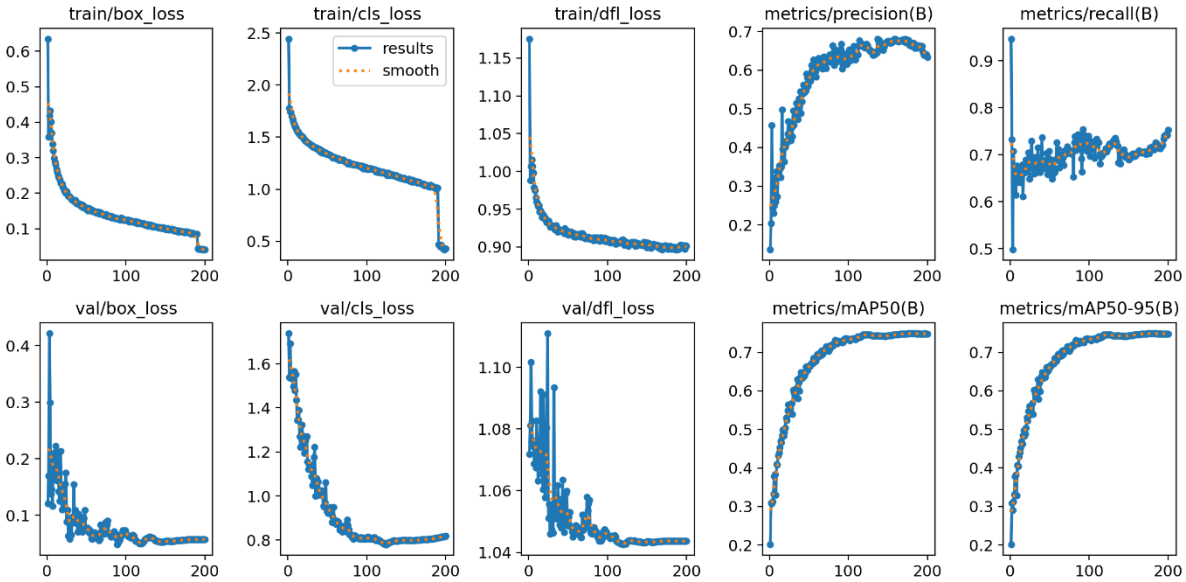
În ansamblu, acest setup urmărește să obțină un model generalizat, capabil să recunoască expresiile faciale în condiții variate, cu un cost rezonabil de calcul.

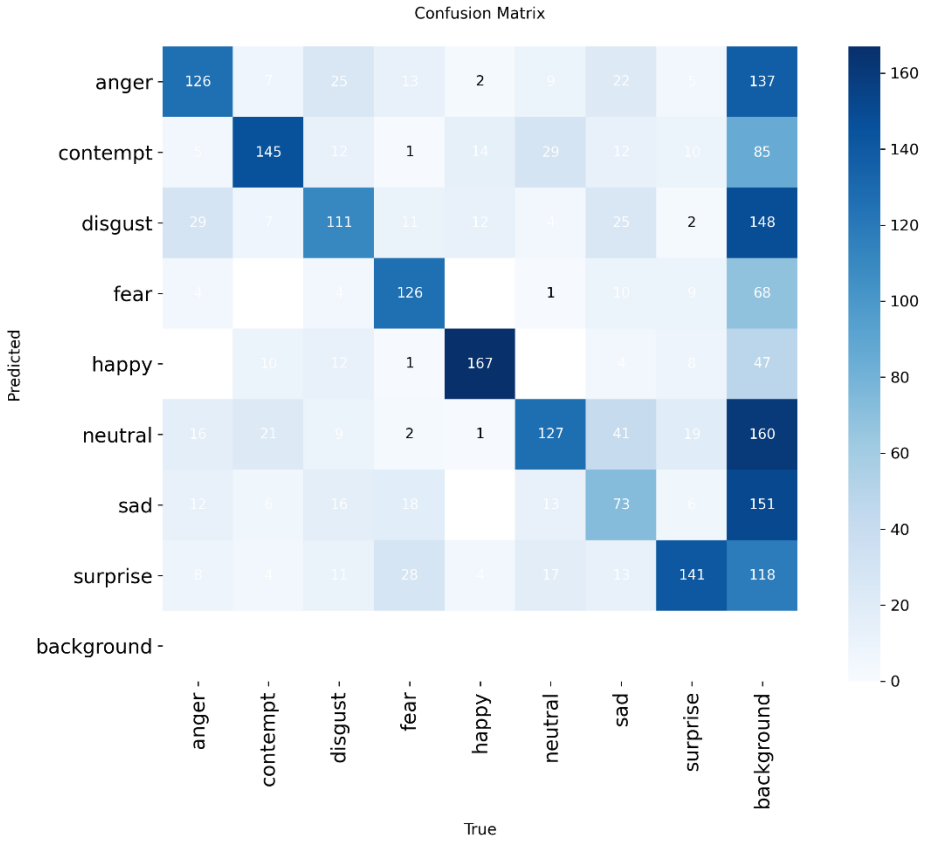
Rezultate obtinute:

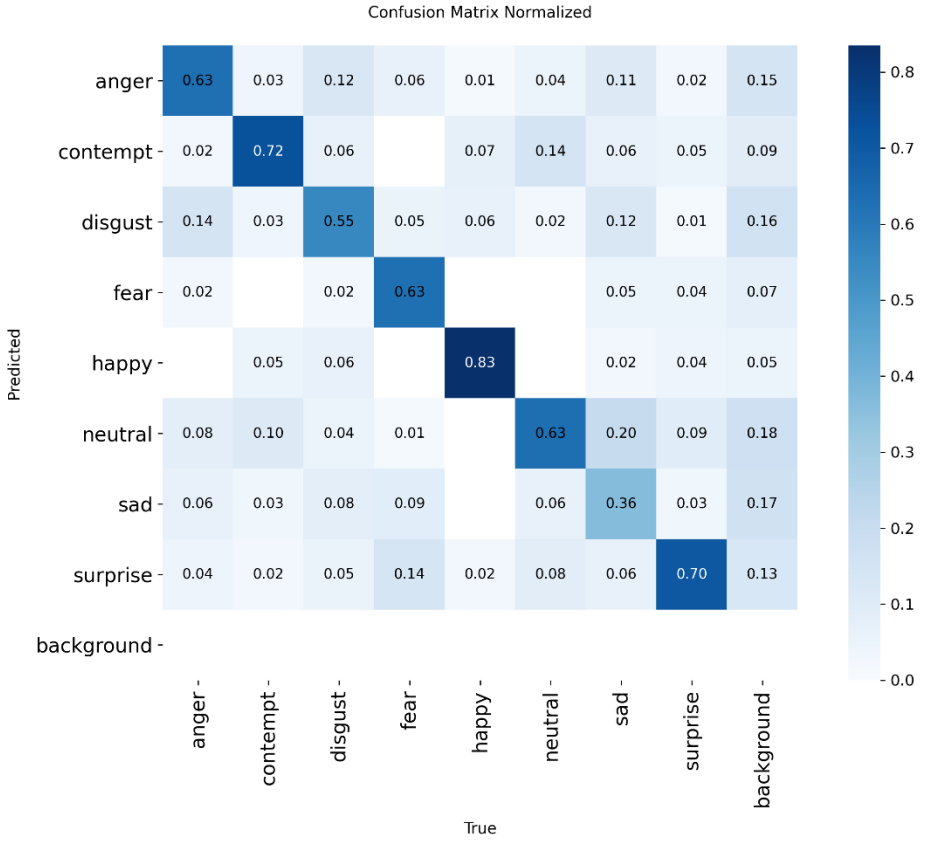


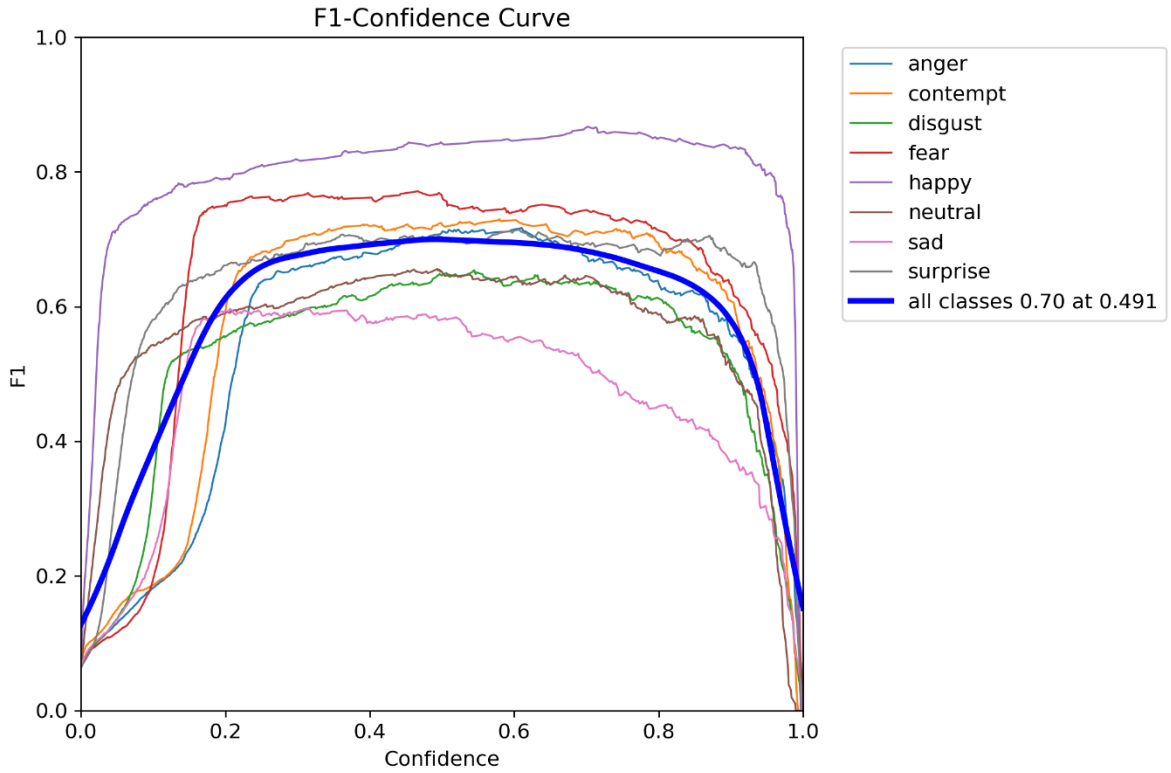



Recall-Confidence Curve
Recall
Confidence
anger
contempt
disgust
fear
happy
neutral
sad
surprise
all classes 0.99 at 0.000


train/box_loss
train/cls_loss
results
smooth
train/dfl_loss
metrics/precision(B)
metrics/recall(B)
val/box_loss
val/cls_loss
val/dfl_loss
metrics/mAP50(B)
metrics/mAP50-95(B)


Confusion Matrix
anger
contempt
disgust
fear
happy
neutral
sad
surprise
background
Predicted
True
160
140
120
100
80
60
40
20
0


Confusion Matrix Normalized
anger
contempt
disgust
fear
happy
neutral
sad
surprise
background
Predicted
True
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0



Astfel pentru fiecare emoție avem:

1. Happy

- **mAP@0.5 = 0.926**, **F1_max ≃ 0.84 (la conf. ≃ 0.45)**
- **Precision–Recall**: Precision începe la ≃ 0.70 chiar și la confidențe mici, apoi urcă spre 0.90–0.95 la confidențe medii (0.4–0.6) și se menține ridicată până la 1.0 la confidențe > 0.8. Recall rămâne peste 0.80 pentru o gamă largă de praguri (0.2–0.8).
- **Confuziuni**: <1% din fețele „happy” au fost clasificate greșit—sporadic cu „surprise” dacă gura e deschisă larg.
- **Motiv**: Zâmbetul deschis (dentiție vizibilă) creează feature-uri consistente pe care rețeaua le detectează foarte bine.

2. Fear

- **mAP@0.5 = 0.796**, **F1_max ≃ 0.77 (la conf. ≃ 0.50)**
- **Precision–Recall**: Precision rămâne peste 0.75 în intervalul confidențelor 0.3–0.7; recall atinge vârfuri de 0.85 la confidențe mici și scade spre 0.65 la valori > 0.7.
- **Confuziuni**: ~10% din „fear” au fost clasificate ca „surprise” (ochi mari + gură deschisă), și ~5% ca „neutral” în cadre mai întunecate.
- **Motiv**: Sprâncenele ridicate și privirea larg deschisă disting bine, dar overlap-ul cu „surprise” e inevitabil.

3. Surprise

- **mAP@0.5 = 0.790**, **F1_max ≃ 0.71 (la conf. ≃ 0.55)**

- **Precision–Recall**: Precision crește rapid peste 0.80 la confidențe ≥ 0.4; recall rămâne > 0.75 până la confidențe de 0.8.
- **Confuziuni**: ~7% „surprise” marcate ca „fear”, și 5% ca „happy” în cazurile de zâmbet larg plus ochi deschiși.
- **Motiv**: Structura feței (guriță + sprâncene) produce semnale clare, dar amestecul cu teamă sau bucurie scade puțin acuratețea.

4. Contempt

- **mAP@0.5 = 0.738**, **F1_max ≃ 0.72 (la conf. ≃ 0.50)**
- **Precision–Recall**: Precision atinge ~0.75–0.80 la confidențe 0.4–0.7, dar recall scade rapid sub 0.60 la praguri > 0.7.
- **Confuziuni**:
    - 12% din „contempt” confundate cu „neutral” (colțul gurii puțin ridicat poate părea față fără expresie).
    - 8% ca „sad” în cadre slab iluminate.
- **Motiv**: Semnalul este foarte subtil—un singur colt al gurii ridicat—iar variațiile legate de unghi și iluminare îi afectează detectarea.

5. Anger

- **mAP@0.5 = 0.712**, **F1_max ≃ 0.71 (la conf. ≃ 0.50)**
- **Precision–Recall**: Precision stabilă la ~0.60 pentru confidențe mici, urcă spre 0.75 la confidențe 0.4–0.6, apoi scade lent. Recall rămâne > 0.80 până la confidențe de 0.6.
- **Confuziuni**:
    - 10% marcate ca „disgust” (sprâncene încruntate + buze strânse).
    - 8% ca „sad” când colțurile gurii sunt ușor coborâte.
- **Motiv**: Furia se manifestă prin trăsături dure (încruntat), dar overlap-ul cu expresii apropiate poate duce la erori.

6. Neutral

- **mAP@0.5 = 0.710**, **F1_max ≃ 0.66 (la conf. ≃ 0.45)**
- **Precision–Recall**: Precision începe foarte jos (~0.40) la praguri scăzute și crește lent spre 0.65–0.70; recall este ridicat (~0.85) la confidențe ≤ 0.3, apoi scade spre 0.60 la praguri ridicate.
- **Confuziuni**:
    - 15% clasificate ca „sad” sau „contempt”.
    - 12% ca „happy” în cadre cu expresii neutre ușor zâmbitoare.
- **Motiv**: Absența expresiei clare face ca neutral să fie „catch-all” pentru multe fețe, rezultând multe false positive.

7. Disgust

- **mAP@0.5 = 0.656**, **F1_max ≃ 0.64 (la conf. ≃ 0.55)**
- **Precision–Recall**: Precision ~0.55–0.65 pentru confidențe 0.3–0.7; recall scade rapid sub 0.50 la praguri > 0.6.
- **Confuziuni**:
    - 14% din „disgust" clasificate ca „anger" (ridicarea nărilor + privire aspră).
    - 10% ca „sad" în imagini cu rezoluție slabă.
- **Motiv**: Micro-expresiile de dezgust sunt subtile și adesea mascate de zgomot de fundal sau rezoluție insuficientă.

8. Sad

- **mAP@0.5 = 0.578**, **F1_max ≃ 0.59 (la conf. ≃ 0.35)**
- **Precision–Recall**: Precision maxim în jur de 0.60 la confidențe 0.3–0.5, apoi scade; recall pornește la ~0.75 la praguri joase și scade sub 0.50 la praguri > 0.6.
- **Confuziuni**:
    - 18% marcate ca „neutral" (colțuri gurii coborâte discret).
    - 12% ca „contempt" în ipostaze cu un singur colț al gurii ridicat.
- **Motiv**: Variabilitatea expresiei triste și asemănarea cu neutral fac din „sad" cel mai dificil de detectat.

Rezumat general al curbelor de confidence

- Curba F1 medie atinge maxim ≃ 0.70 la prag ≃ 0.49.
- Precision–confidence mediană ≃ 0.62 la prag ≃ 0.52; recall–confidence median ≃ 0.78 la prag ≃ 0.30.
- Alegerea pragului optim (≈ 0.5) este un bun compromis pentru majoritatea claselor, dar pentru aplicații specifice se poate ajusta (de ex. prag mai jos pentru recall ridicat pe „fear" și „sad").

Recomandări punctuale

- Pentru clasele subtile (*contempt*, *neutral*, *sad*, *disgust*), adăugarea de augmentări focalizate pe aceste expresii (crop-uri care pun în evidență zona gurii/nărilor) și utilizarea focal loss pot îmbunătăți recall-ul și precision-ul.
- Pentru aplicații critice unde recall-ul e prioritar (ex. detecție „fear" în siguranță), setați pragul de încredere între 0.3–0.4. Pentru precision strict (ex. marketing „happy"), pragul poate fi mărit la 0.6–0.7.

Aceste detalii oferă o privire completă asupra comportamentului modelului clasa cu clasa și te ajută să decizi ajustările viitoare.

Etapa de verificare pe exemple (Vizualizare batch-uri)

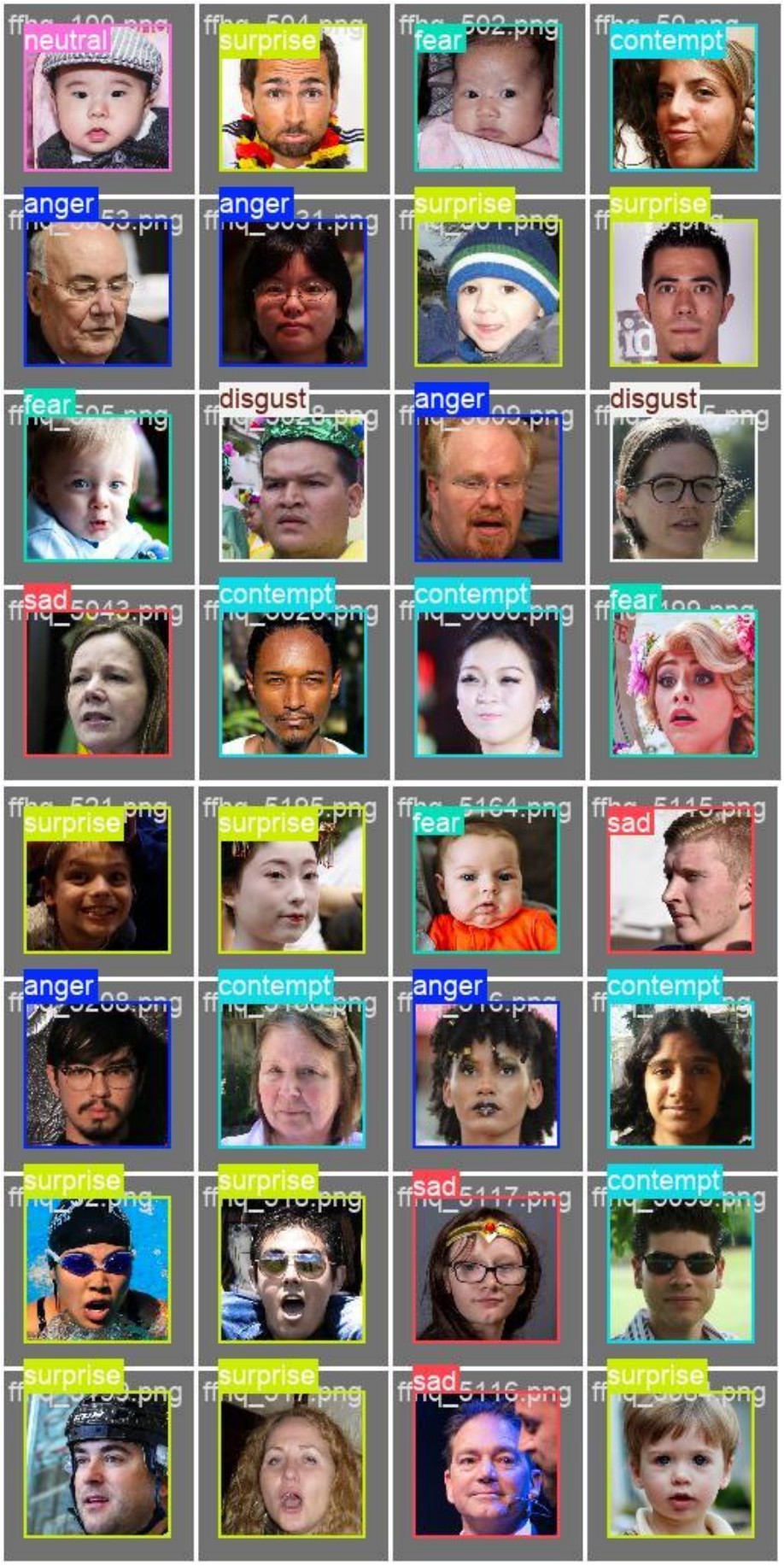
Pentru a verifica calitatea etichetării și a augmentărilor, am extras batch-uri aleatorii din setul de testare și am suprapus:

- **Cutii de delimitare (bounding boxes)** color-code după ID-ul clasei.
- **Numărul clasei** afișat în colțul fiecărei cutii.
- **Imagini redimensionate** la 96×96 px.

Interpretare:

- Fiecare imagine din batch reprezintă fețe în diferite stări și condiții de iluminare.
- Culorile boxelor permit identificarea rapidă a claselor (ex: roșu pentru Disgust, albastru pentru Neutral, verde pentru Sad etc.).
- Numerele confirmă etichetele corespunzătoare fiecărei expresii.
- Această vizualizare ajută la detectarea erorilor de etichetare sau a cazurilor dificile (ex: fețe rotite sau expresii ambigue).
- Fiecare sentiment are un anumit număr asociat:

```
0: anger
1: contempt
2: disgust
3: fear
4: happy
5: neutral
6: sad
7: surprise
```


neutral
surprise
fear
contempt
anger
anger
surprise
surprise
fear
disgust
anger
disgust
sad
contempt
contempt
fear
surprise
surprise
fear
sad
anger
contempt
anger
contempt
surprise
surprise
sad
contempt
surprise
surprise
sad
surprise