Outpainting Localizer

Multimedia Data Security course project

Matteo Darra (6 CFU)
Leonardo Vicentini (3 CFU)

Friday 20th October, 2023

Project purpose

The goals of the project were:

1) **Train** and **test** a model for outpainting localization

2) **Improve metrics** from the obtained baseline

Task and Dataset

- Given a sequence of frame (video) in input, the model goal is to localize which areas of the frames are outpainted

- The dataset is composed of frames and relative binary masks that indicate whether or not each pixel of the frame is outpainted

Model

- **RAFT** (Recurrent All-Pairs Field Transforms) is a model used to predict the **optical flow** of a video

Optical flow

"Optical flow is the task of estimating per-pixel motion between video frames. It is a long-standing vision problem that remains unsolved. The best systems are limited by difficulties including fast-moving objects, occlusions, motion blur, and textureless surfaces". [2]

- In this setting, RAFT is leveraged to do localization of the outpainted region

RAFT architecture

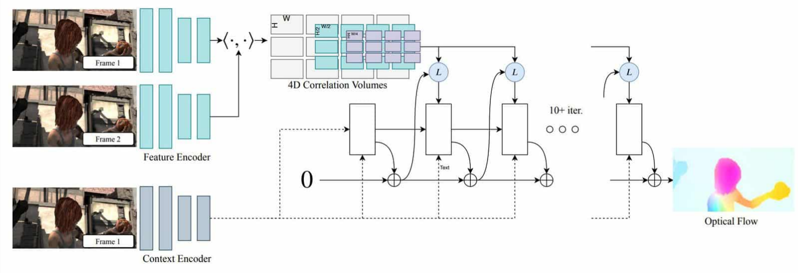

Figure 1: RAFT architecture

- Feature encoder
- Context encoder
- Update module

- Usage logic: video manipulations (outpainting) generate spatial and temporal inconsistencies, detectable using optical flow

Setup

- Become familiar with the task and codebase
- Environment and code setup
  - Test script adaptation
- Check dataset integrity: identified empty folders

Metric

- F1_SCORE is a measure of a test's accuracy
- F1_SCORE formula:

$$F1\_SCORE = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

where:

- **precision** is the number of true positive results divided by the number of all positive results, including those not identified correctly
- **recall** is the number of true positive results divided by the number of all samples that should have been identified as positive

Baseline

- Baseline was obtained by training the model as-is
  - Loss function: weighted softmax cross entropy
- F1_SCORE: 0.501
- Variance: 0.011

Baseline examples (good performance)

Figure 2: Frame
Video type: outdoor, move;
F1_score: 0.921

Figure 3: Generated mask

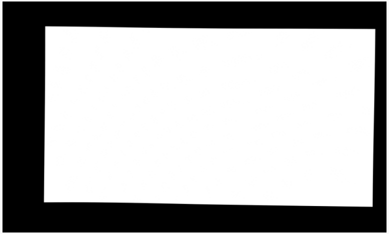

Figure 4: Ground truth mask

Baseline examples (poor performance)

Figure 5: Frame
Video type: outdoor, move;
F1_score: 0.050

Figure 6: Generated mask

Figure 7: Ground truth mask

Data Augmentation

- Started with high-level intuitions before the actual implementations
- Attempted to devise augmentations that simulated real variations that might be encountered during video frame capture
- Goal: improve model's **generalization** capabilities
- Strategies used:
  - Horizontal flipping
  - Time warping
  - Random elastic deformation

Rationale

- Samples doubling
- Camera movement could be biased toward a specific side, e.g. from left to right
- Vertical flipping: unrealistic case, not implemented

Figure 8: Original

Figure 9: Flipped

Horizontal flipping

- Results:
  - F1_SCORE: 0.627
  - Variance: 0.016
- 25% improvement from baseline

Horizontal flipping only examples (good performance)

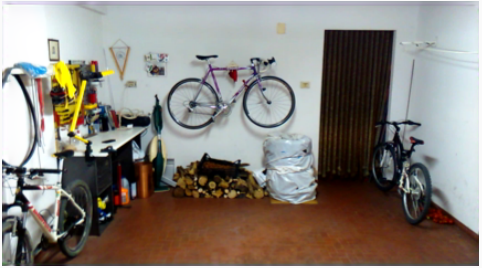

Figure 10: Frame
Video type: indoor, still;
F1_score: 0.953

Figure 11: Generated mask

Figure 12: Ground truth mask

Horizontal flipping only examples (poor performance)

Figure 13: Frame
Video type: indoor, panrot;
F1_score: 0.121

Figure 14: Generated mask

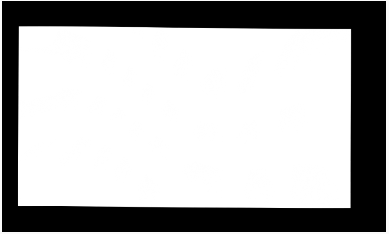

Figure 15: Ground truth mask

Focal loss formula

$$\text{FocalLoss} = -\frac{1}{N}\sum_{i=1}^{N}\left(\alpha \cdot \left(1 - \exp\left(-\text{BCE}(\text{flow\_preds}[i], \text{masks}, \text{weight})\right)\right)^{\gamma}\right.$$
$$\left. \cdot \text{BCE}(\text{flow\_preds}[i], \text{masks}, \text{weight})\right)$$

- It is a function similar to default loss function used (weighted softmax cross entropy loss)
- It achieved slightly **worse results**

Gamma	F1_SCORE	Variance
0.7	0.603	0.014
0.8	0.610	0.011
0.9	0.568	0.016

Table 1: F1_SCORE and variance with focal loss

Time warping

Rationale

- Swap frame n and $n-1$ to simulate different camera movement

- Results:
  - F1_SCORE: 0.626
  - Variance: 0.011
- Note: time warping is added on top of the horizontal flipping augmentation with a frequency of 1/10
- No improvement from horizontal flipping but with lower variance

Random elastic deformation

Rationale

- To increase the variety of mask shapes to simulate different cameras
- To simulate slight imperfections

- Implementation leveraging **TorchIO** library [1]
- Parameters tuned:
  - max_displacement (MD)
  - num_control_points (CP)
- The same transformation is saved and applied to frame n, frame $n+1$, mask n, mask $n+1$

Random elastic deformation

Figure 16: Original

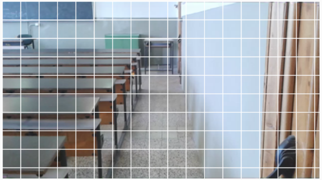

Figure 17: Original with grid

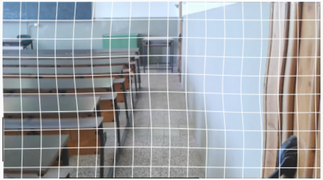

Figure 18: MD: (15,10); CP:7

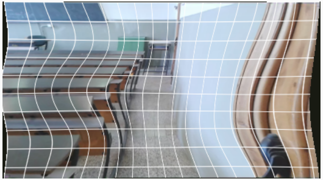

Figure 19: MD: (150, 10); CP:7

Figure 20: MD: (15,10); CP:40

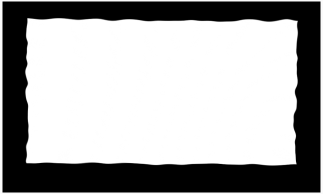

Figure 21: Deformed Mask

Random elastic deformation

Frequency	max_displacement	num_control_points	F1_SCORE	Variance
1/9	15,15	7	0.626	0.009
1/6	18,18	9	0.702	0.010
1/6	**20,20**	**9**	**0.713**	**0.010**
1/4	21,21	11	0.634	0.006

Table 2: F1_SCORE and variance as chosen parameters change (from light to heavy)

- Frequency: how frequent a random elastic deformation is applied while training the model
- Actual configuration: horizontal flipping + time warping + random elastic deformation

Best result

- Global F1_SCORE: 0.713
- Variance: 0.010
- Best: 0.906
- Worst: 0.439
- F1_SCORE metric **improved by 42% from baseline**

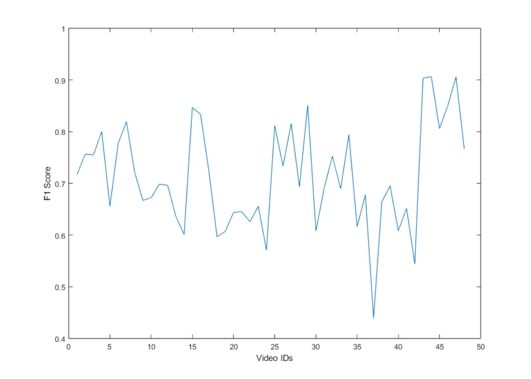

Figure 22: Mean F1_score per video

Best model examples (good performance)

Figure 23: Frame
Video type: outdoor, panrot;
F1_score: 0.957

Figure 24: Generated mask

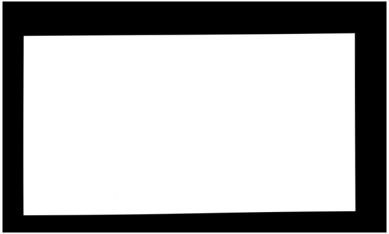

Figure 25: Ground truth mask

Best model examples (poor performance)

Figure 26: Frame

Video type: indoor, move;

F1_score: 0.230

Figure 27: Generated mask

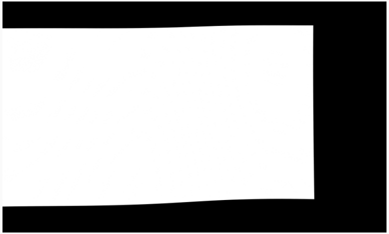

Figure 28: Ground truth mask

Best result

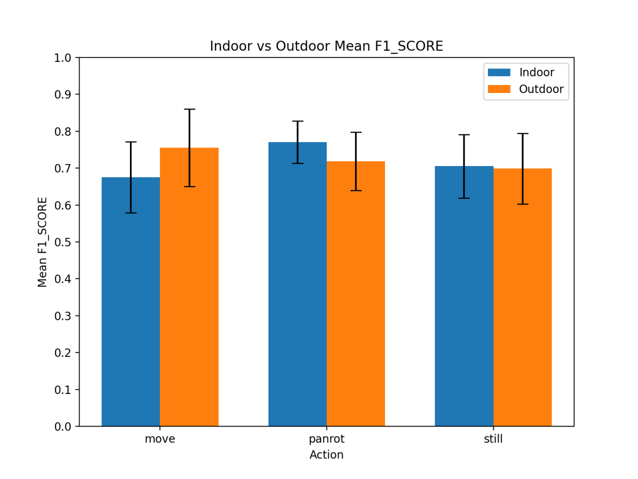

Figure 29: Mean F1_score by video category

Baseline vs. best comparison

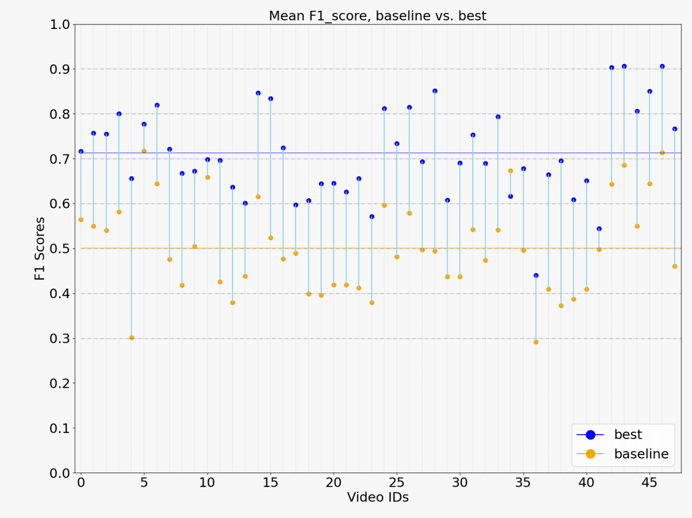

Figure 30: F1_score baseline vs. best, video per video

Baseline vs. best comparison

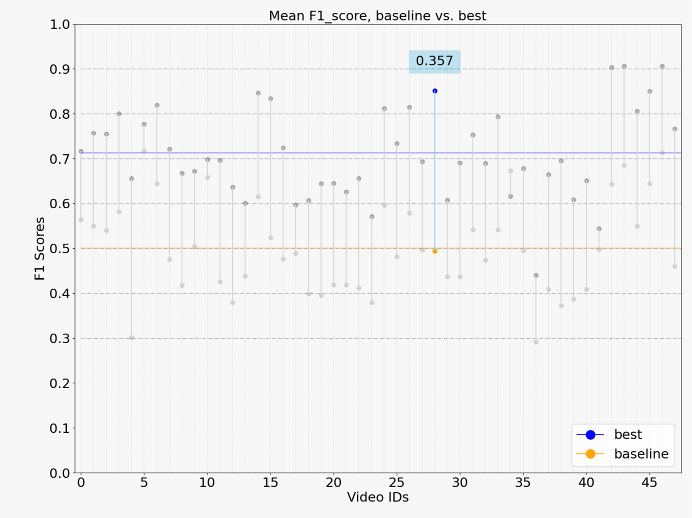

Figure 31: F1_score baseline vs. best, max improvement

Baseline vs. best comparison

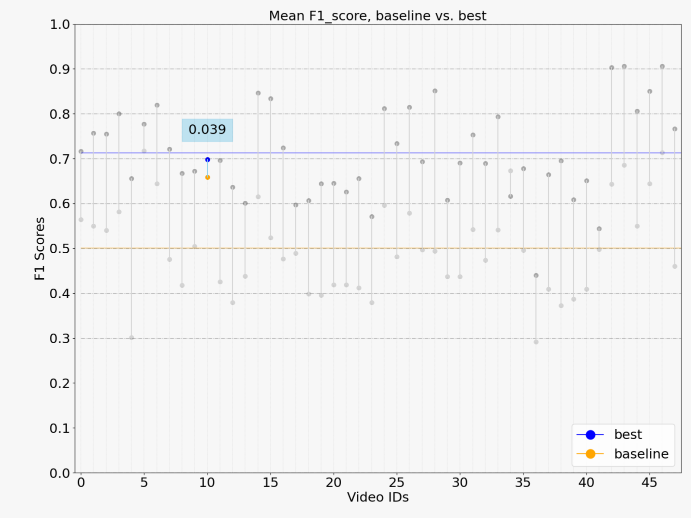

Figure 32: F1_score baseline vs. best, min improvement

Baseline vs. best comparison

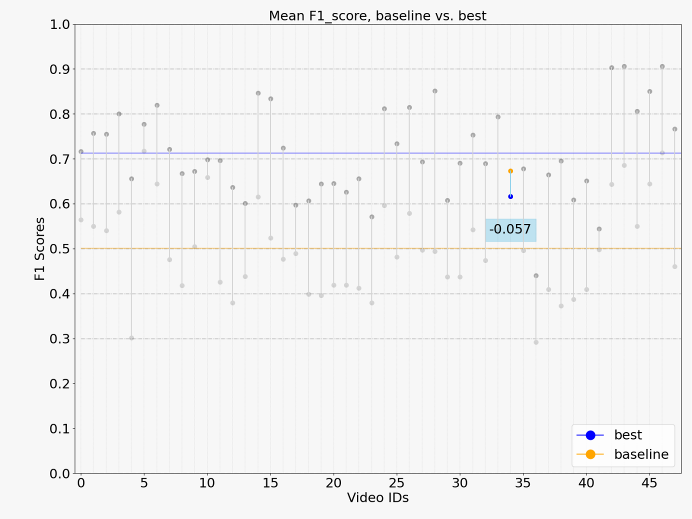

Figure 33: F1_score baseline vs. best, score worsened

Baseline vs. best

	F1_SCORE	Loss	γ	LR	BS	Horizontal Flipping	Time Warping		Random Elastic Deformation			
						ON/OFF	ON/OFF	F	ON/OFF	F	CP	MD
Baseline	0.501	WSCE	0.85	0.0001	12	OFF	OFF	-	OFF	-	-	-
Best	**0.713**	**WSCE**	**0.85**	**0.0001**	**12**	**ON**	**ON**	**1/10**	**ON**	**1/6**	**9**	**20,20**

Mask creation

- Binary masks are built from optical flow using a threshold on a [0,1] scale
- Threshold: 0.5
  - Effect: identified non-outpainted areas are quite small, correspond to the "stronger" areas of optical flow

Figure 34: Mask, varying threshold*

Figure 35: Ground truth mask

- Thresholds: $1 \leq t \leq 0.5$; $0.5 < t \leq 0.4$; $0.4 < t \leq 0.3$; $0.3 < t \leq 0.2$

Mask creation

- Majority of ground truth masks are characterized by large non-outpainted areas
- New threshold: 0.4
  - Effect: identified non-outpainted areas are bigger, containing also less strong areas of optical flow
- Global F1_score: 0.786

Figure 36: Generated mask

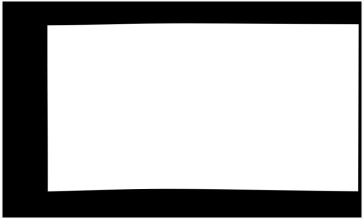

Figure 37: Ground truth mask

Personal contribution

- The team almost always worked together, from ideas brainstorming to implementation and testing

Further improvements

- More thorough parameters exploration
- Other data augmentations methods
- Model architecture modifications
  - Update module adaptation: to exploit temporal information

References

Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin.
Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning.
Computer Methods and Programs in Biomedicine, page 106236, 2021.

Zachary Teed and Jia Deng.
Raft: Recurrent all-pairs field transforms for optical flow.

Thank you for the attention

Figure 38: Baseline

Figure 39: HF

Figure 40: Best