Floor count from street view imagery using learning-based façade parsing

Thesis presentation MSc Geomatics

by Daniël Dobson

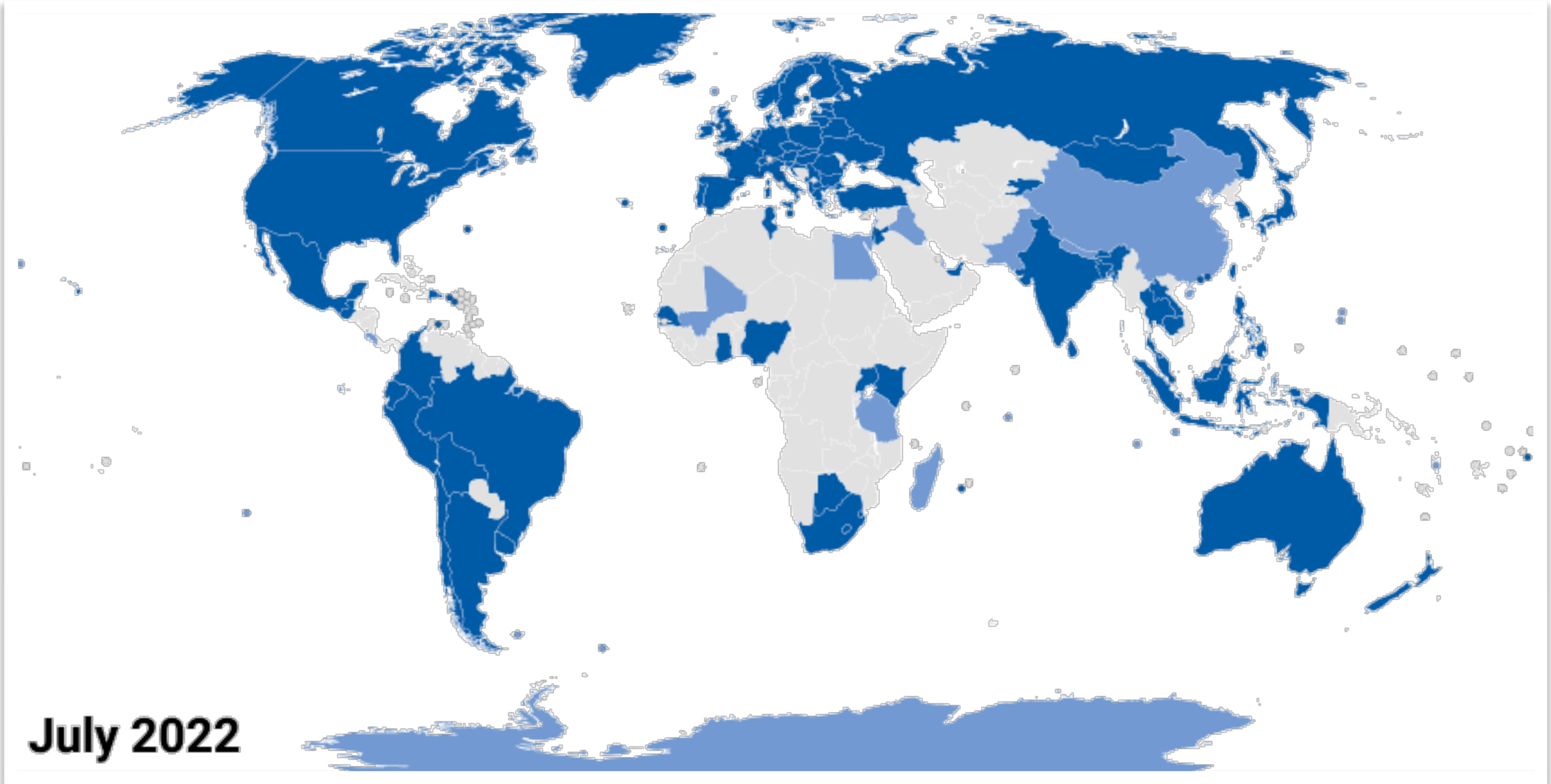
January 20th, 2023

Supervisors: Ken Arroyo Ohori
Nail Ibrahimli
Co-reader: Hugo Ledoux

TUDelft

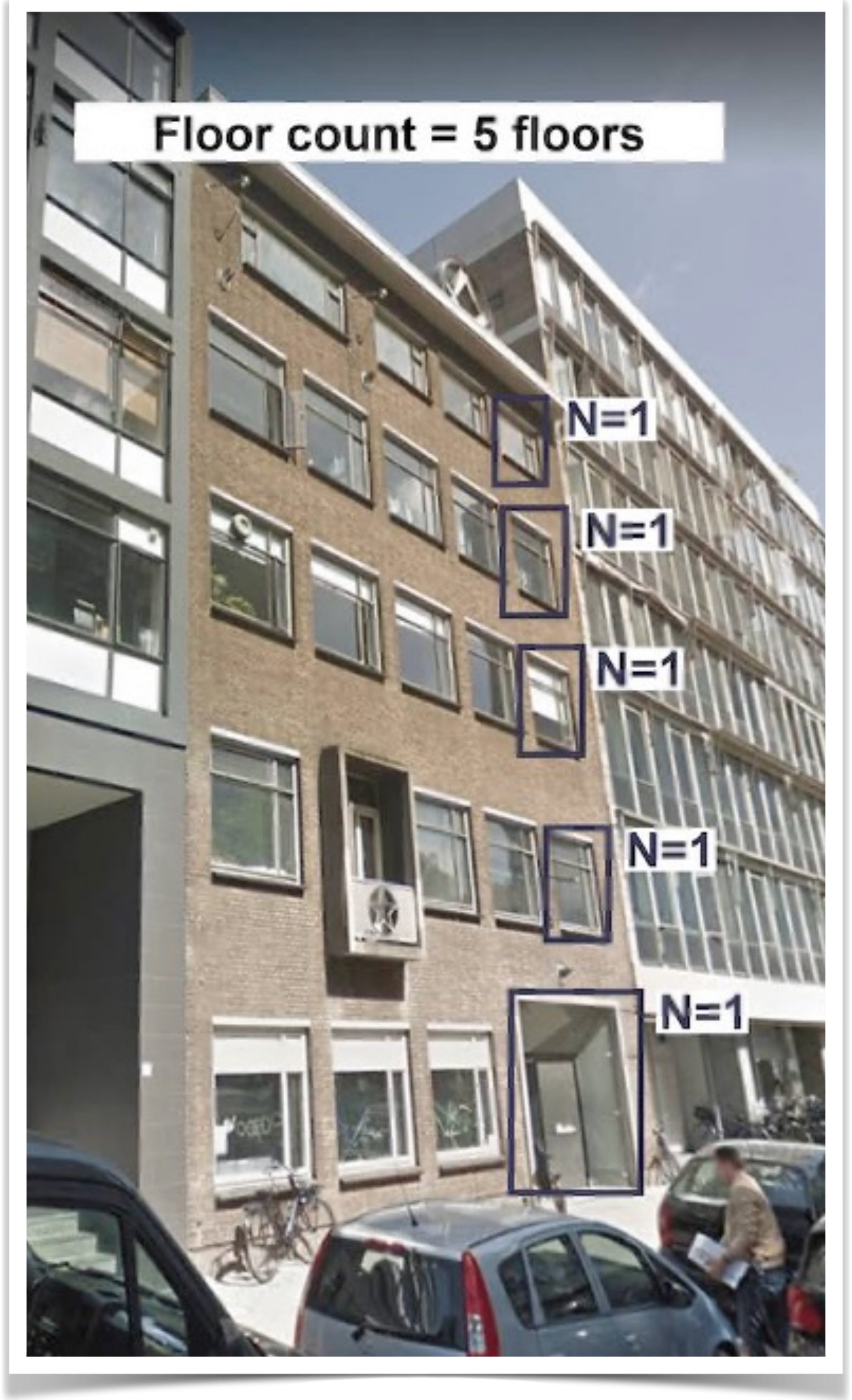
- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

Motivation



[1]

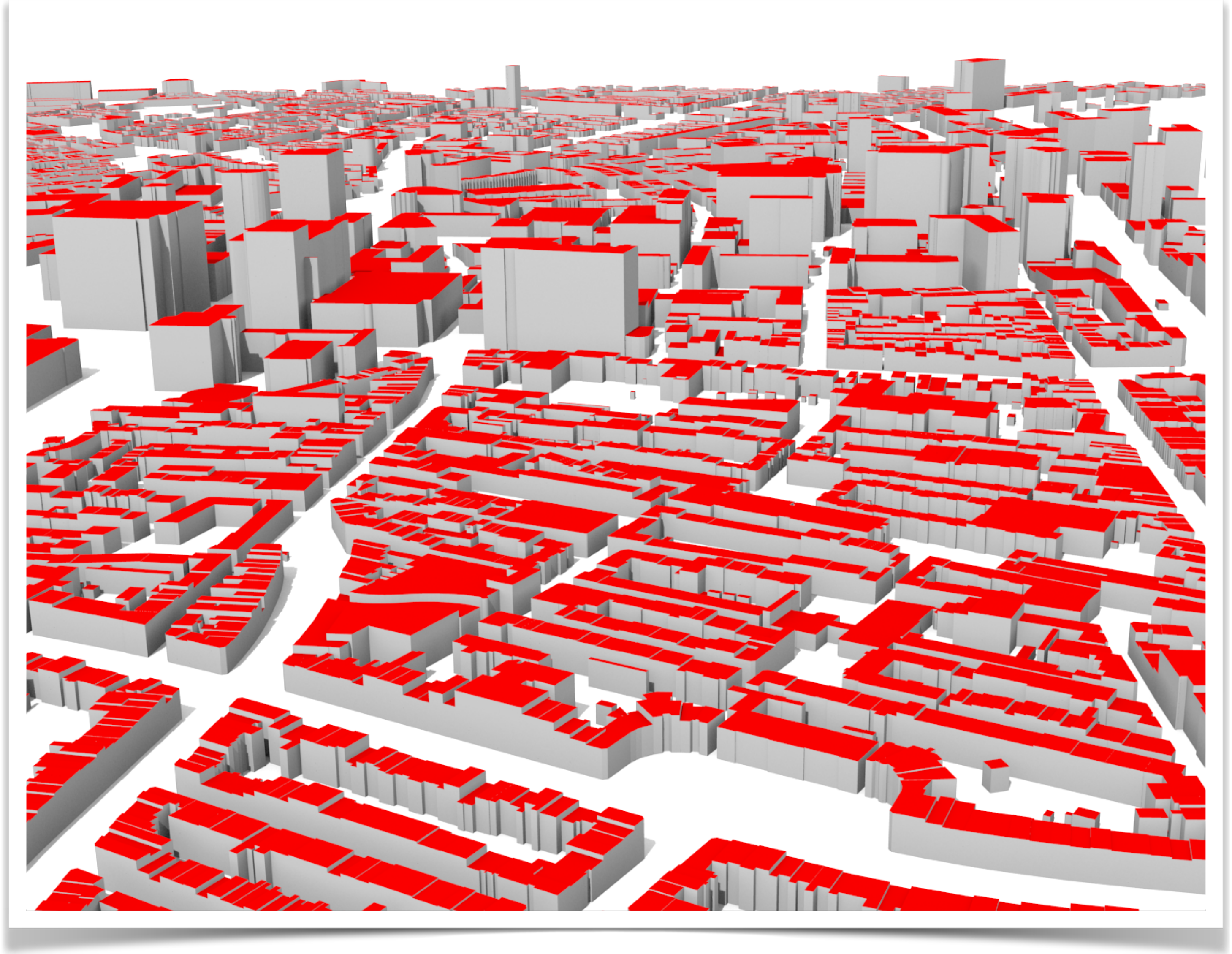
Street View Imagery (SVI)



Façade parsing

- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

Use cases



[2]

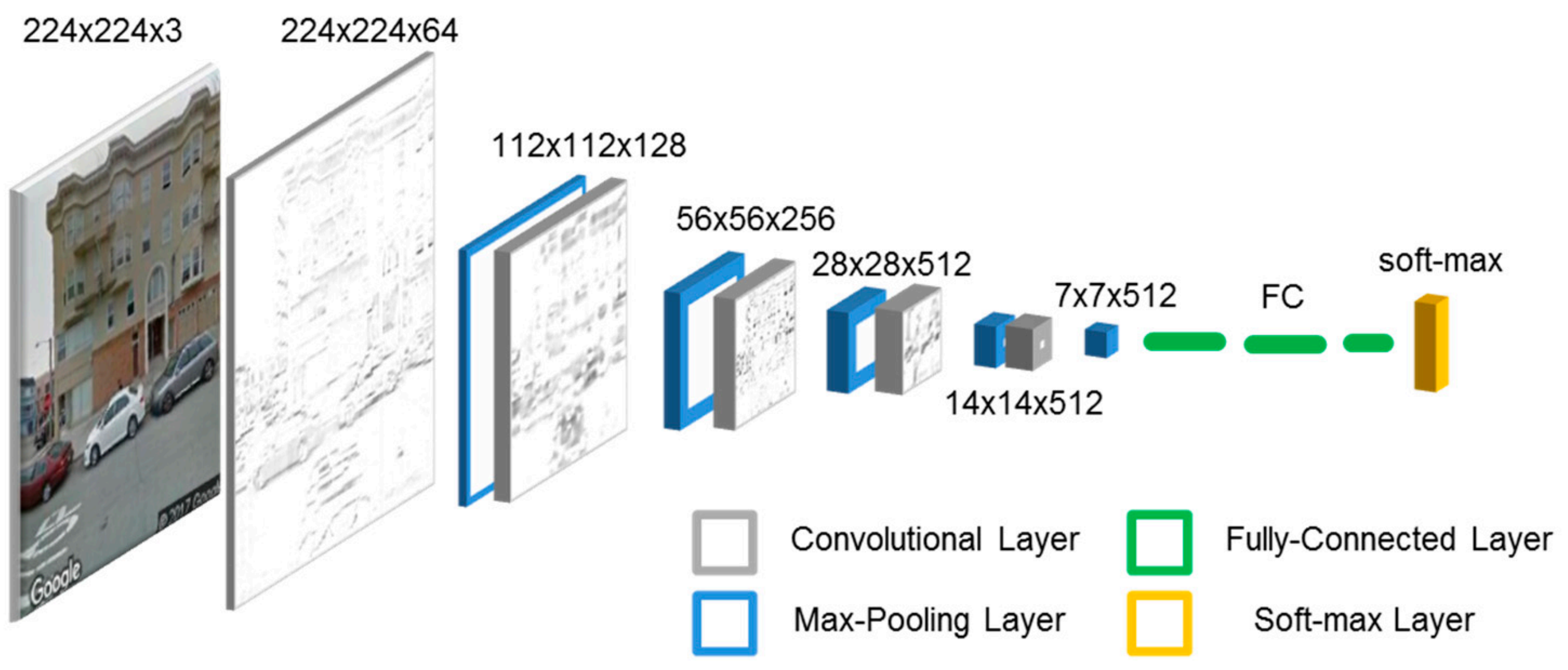
3D city models without elevation data

[2]

Noise pollution modeling

- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

State of the art



[3]

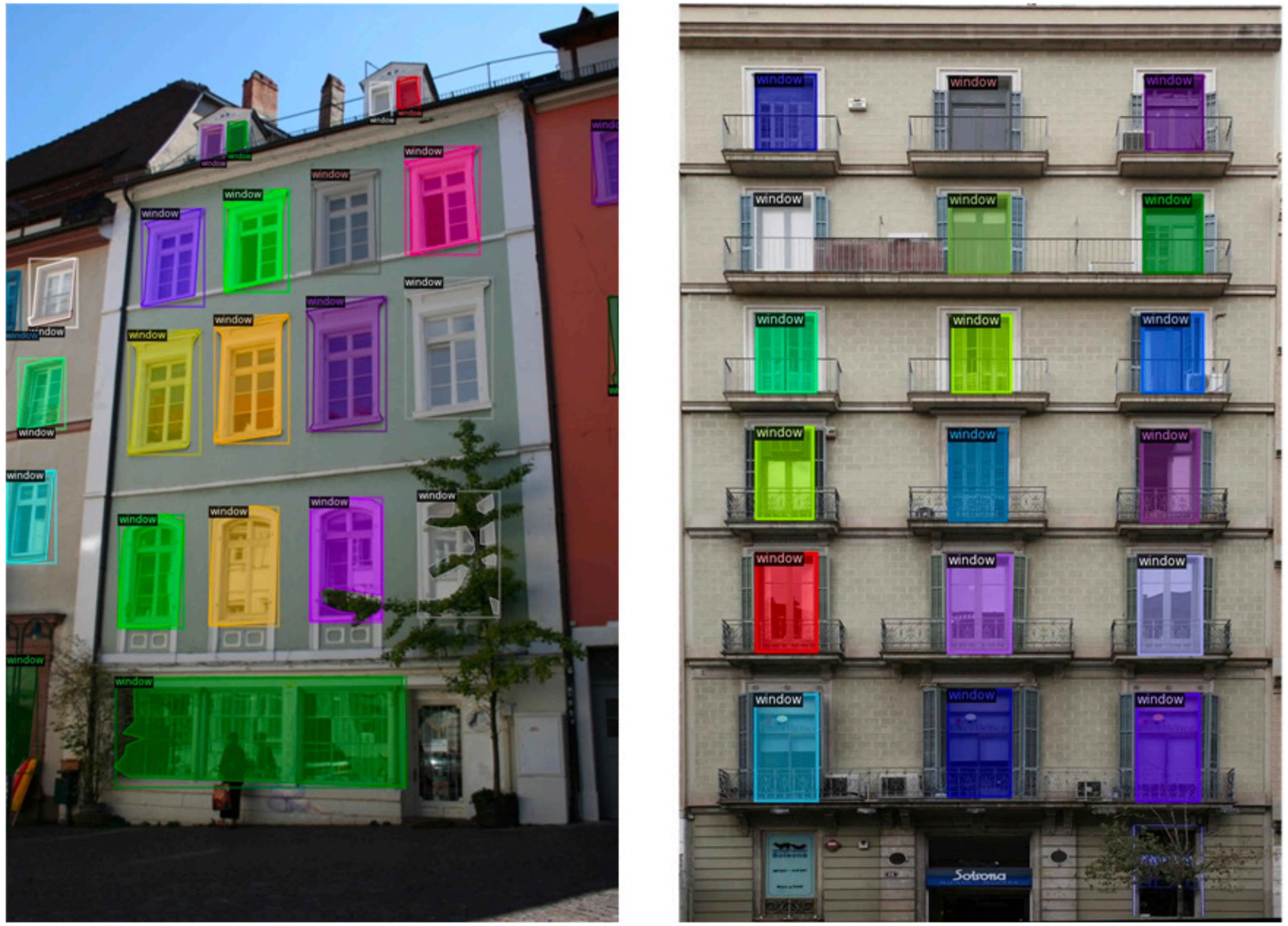
Architecture	Classes	Pre-trained	Accuracy (%)	Train/test images
VGG-16 [3]	0, 1, 2, 3, 4+	✓	85	600/430
ResNet-34 [4]	1, 2, 3	✓	90.5	843/22,803
TREncNet [5]	1, 2+	✓	93.5	33,822/8,593

Limitations:
1. Predefined classes
2. Datasets (bias/size)
3. Unclear learning

Background

- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

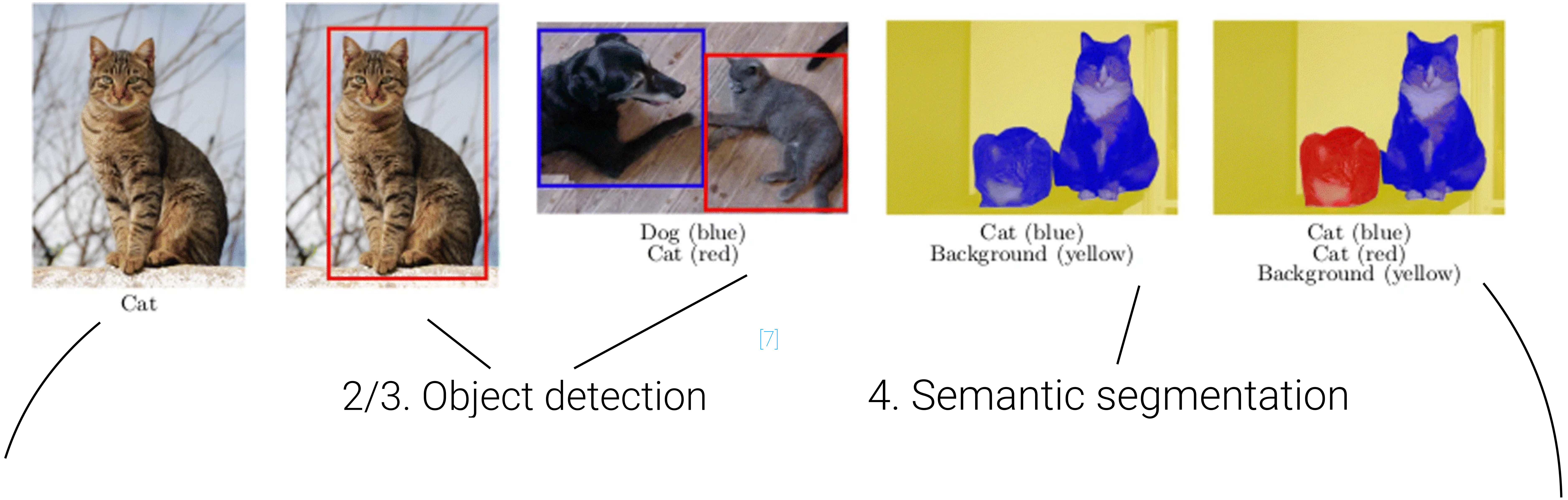
Façade parsing?



[6]

- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

Computer Vision & Deep Learning



[7]

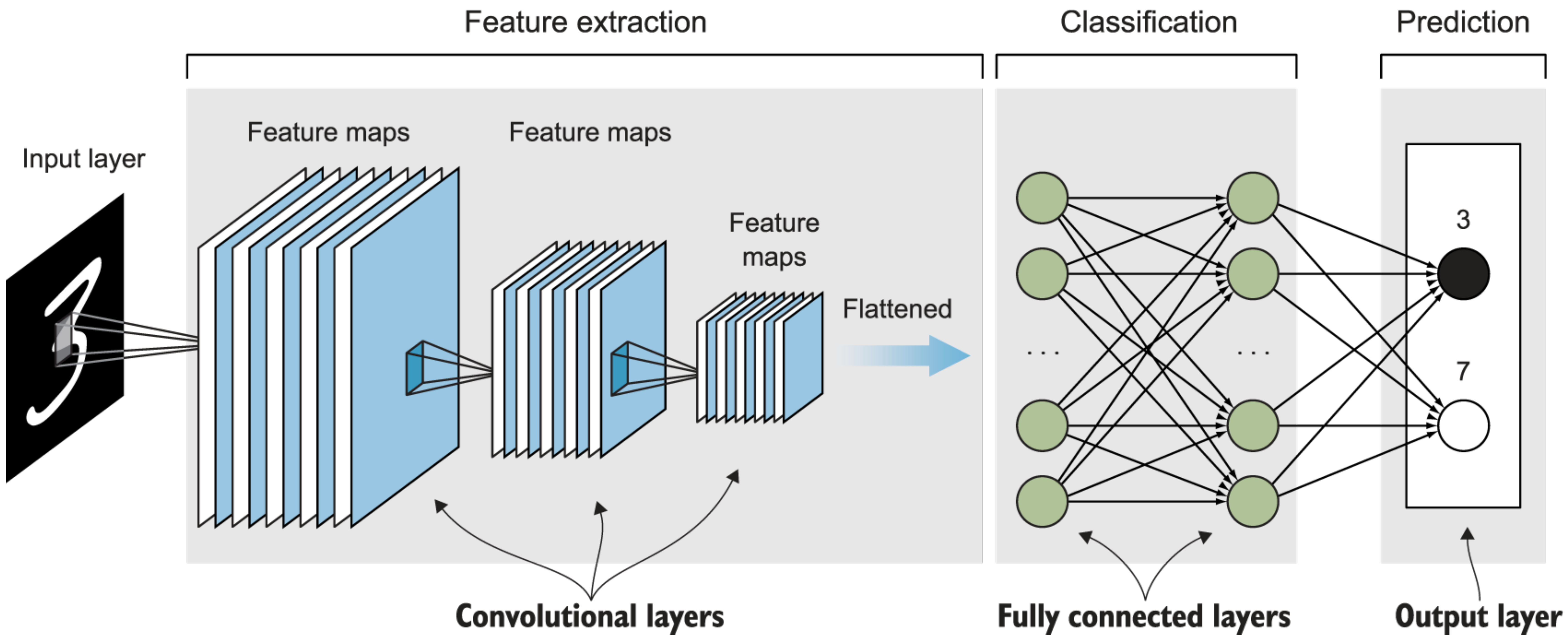
1. Classification

2/3. Object detection

4. Semantic segmentation

5. Instance segmentation

Computer Vision & Deep Learning



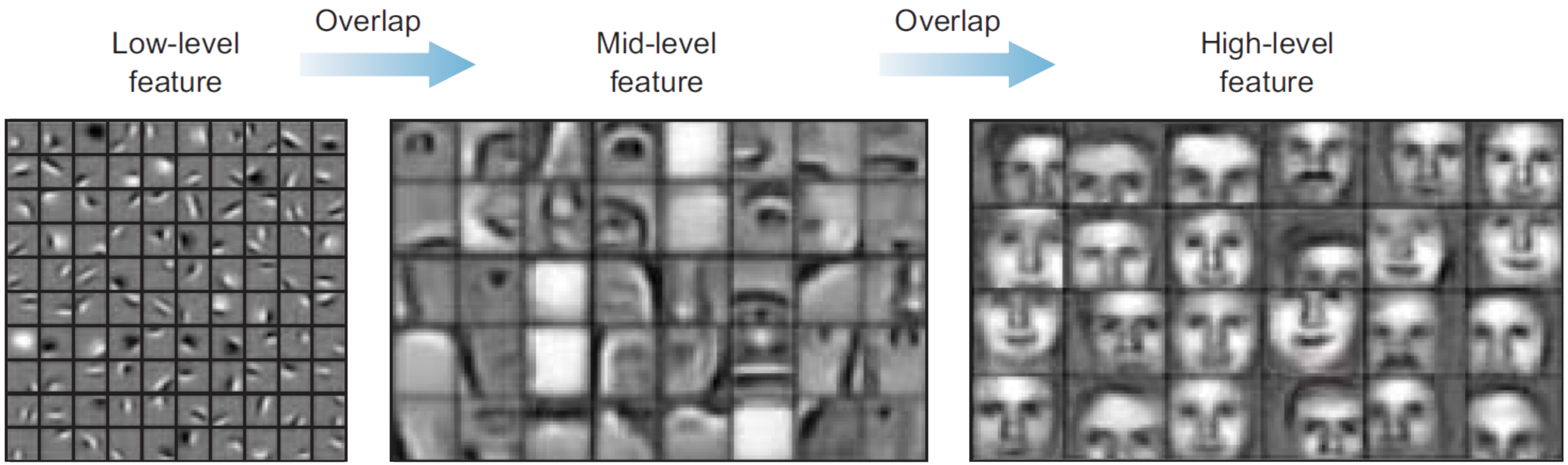
[8]

Convolutional Neural Networks (CNNs)

- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

TUDelft

Computer Vision & Deep Learning



Low-level feature

Overlap

Mid-level feature

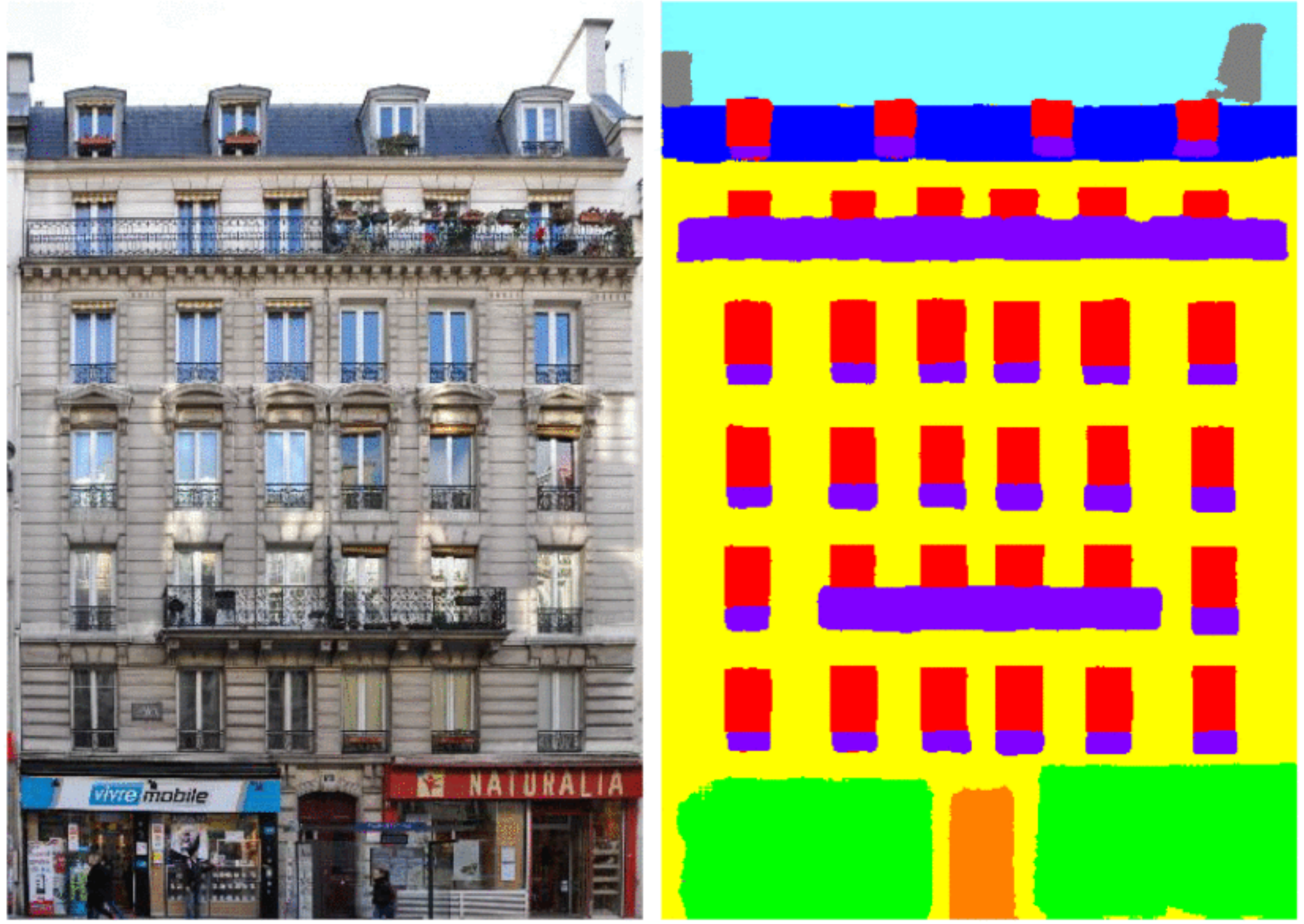
Overlap

High-level feature

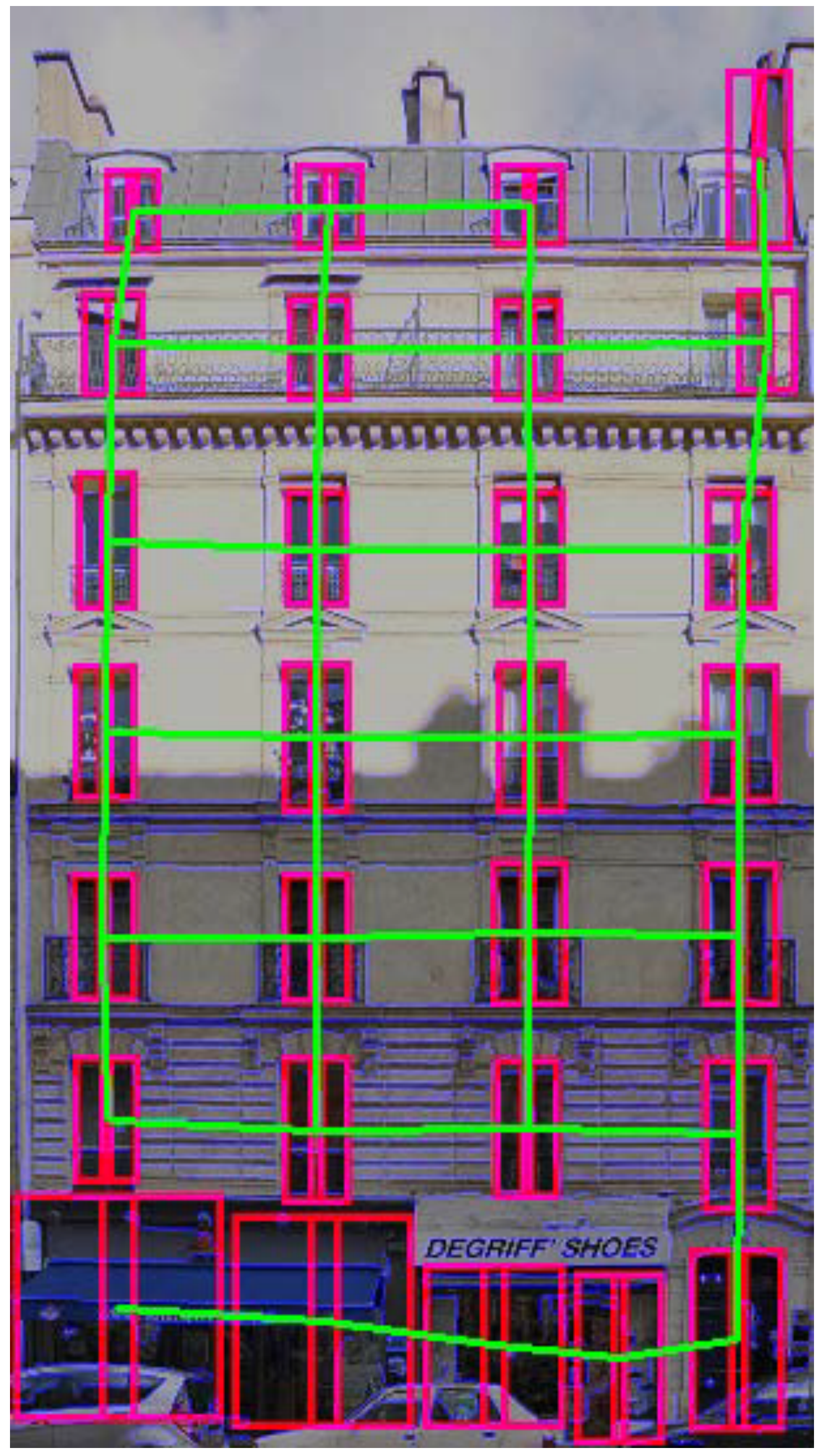
[8]

- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

Façade parsing & Regularity



DeepFacade [9]



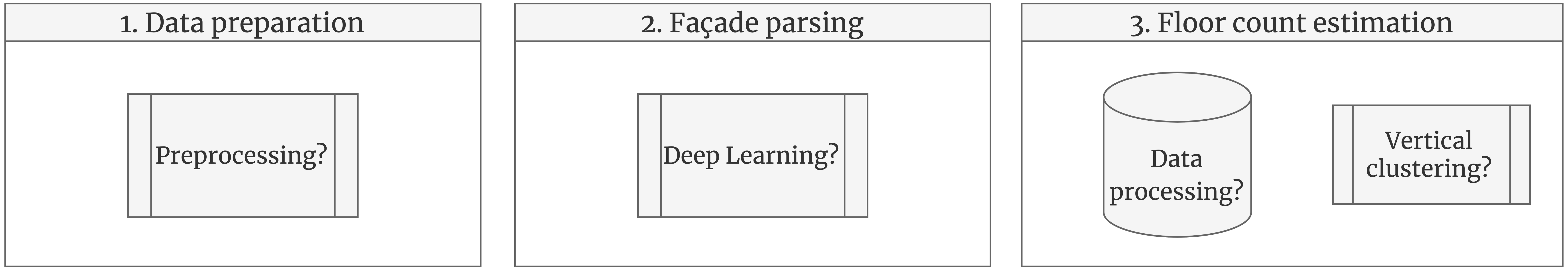
Regularity [10]

Objectives

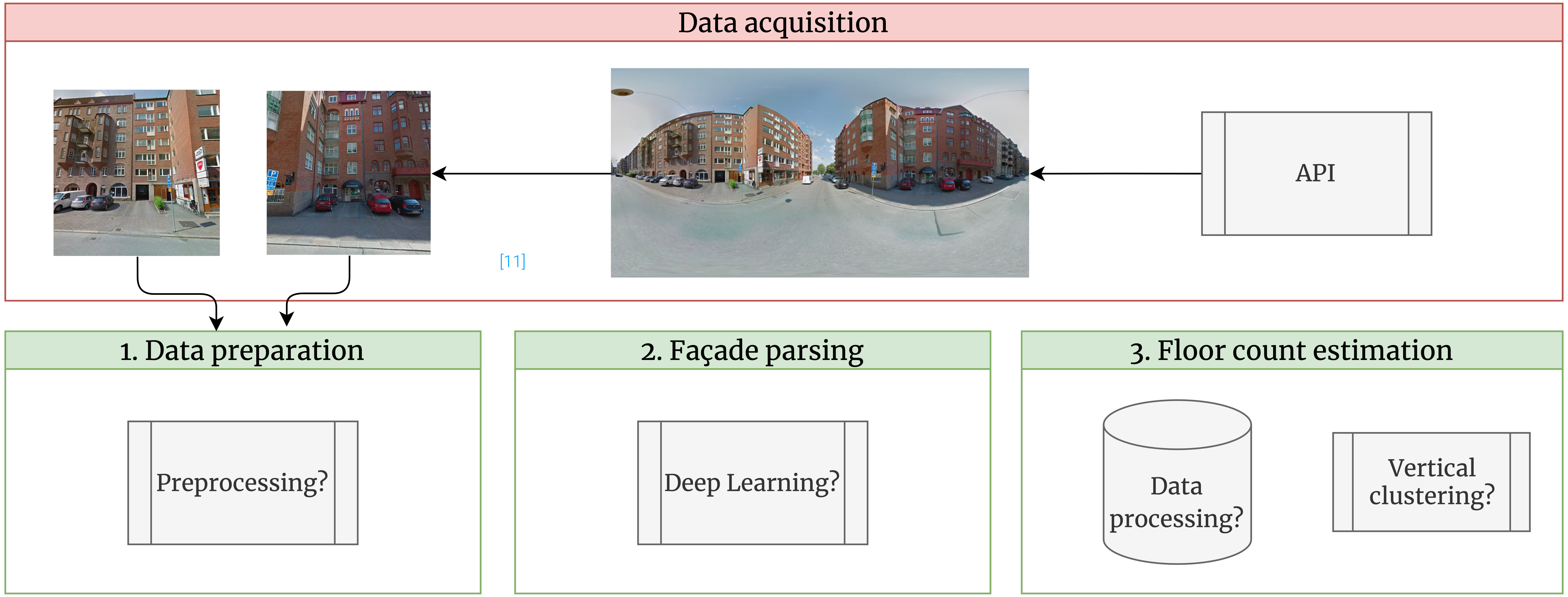
- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

Research questions

How to determine floor count in an image with the use of learning-based façade parsing?



Scope



Data acquisition

API

[11]

1. Data preparation

Preprocessing?

2. Façade parsing

Deep Learning?

3. Floor count estimation

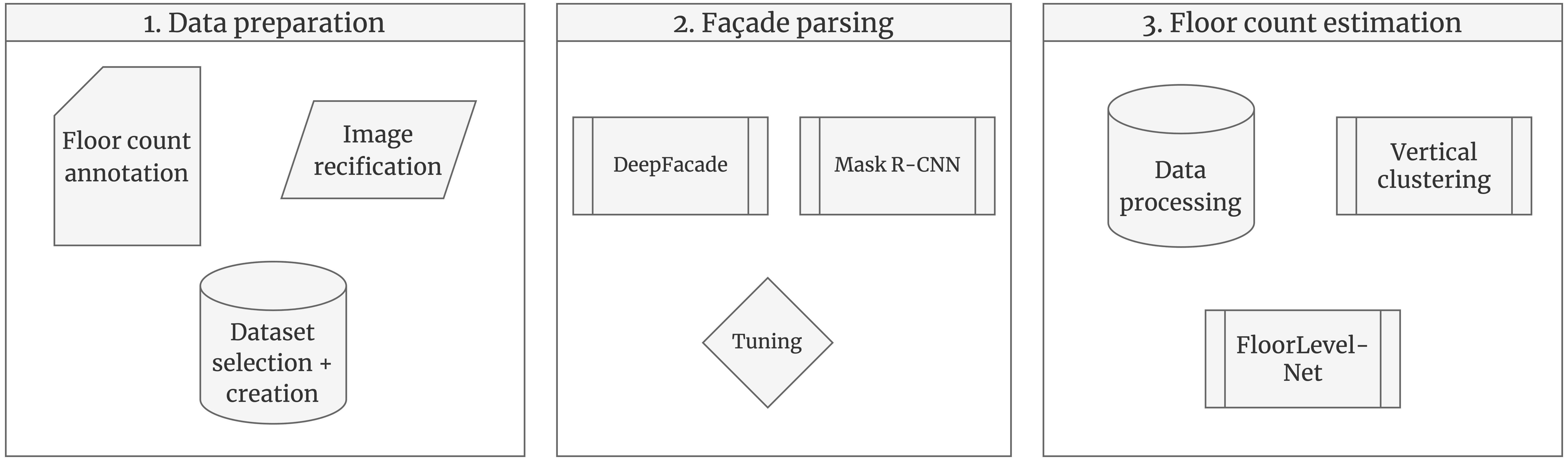
Data processing?

Vertical clustering?

Methodology

- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

Methodology

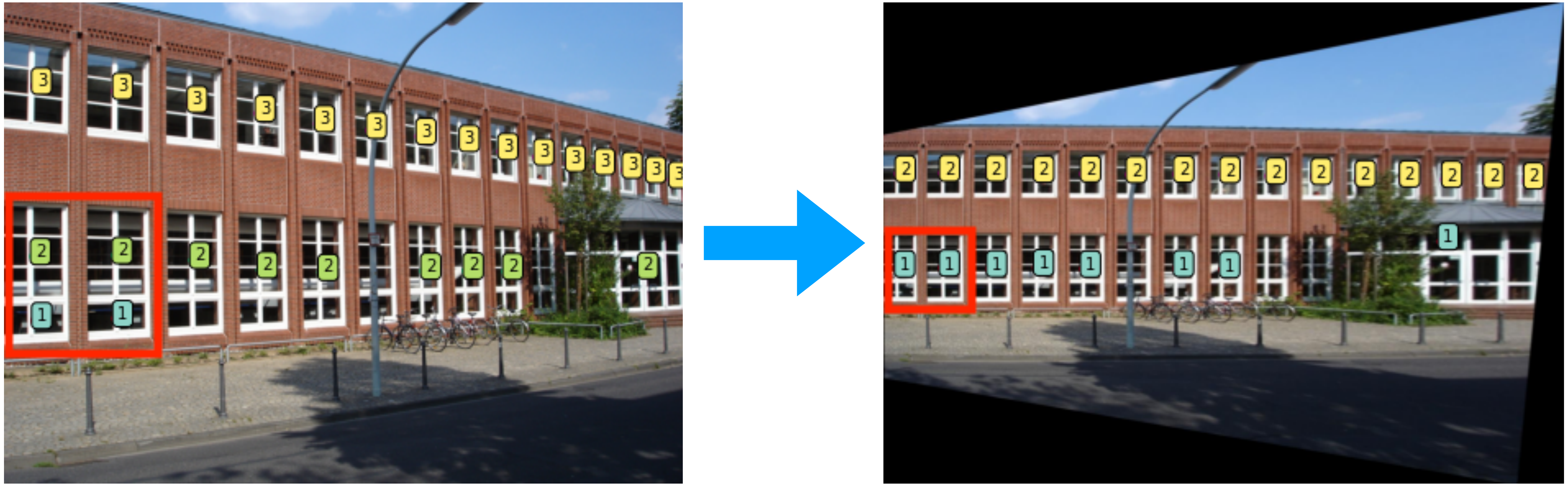


Experiments & Development

- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

TU Delft

Image Rectification

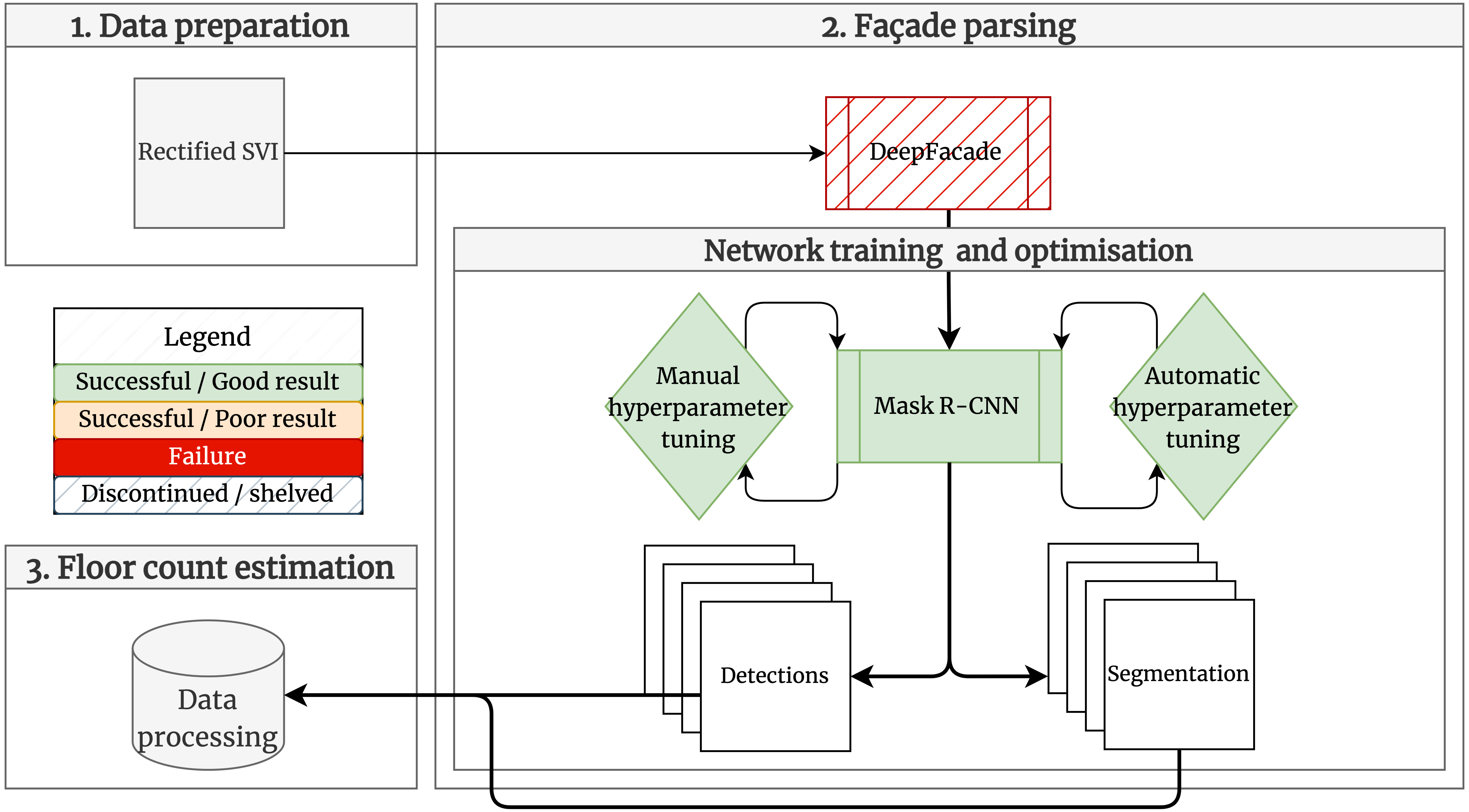


Methods tested:
1. VP estimation
2. Direct homography transform

- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

TUDelft

Façade parsing



1. Data preparation

Rectified SVI

2. Façade parsing

DeepFacade

Network training and optimisation

Manual hyperparameter tuning

Mask R-CNN

Automatic hyperparameter tuning

Detections

Segmentation

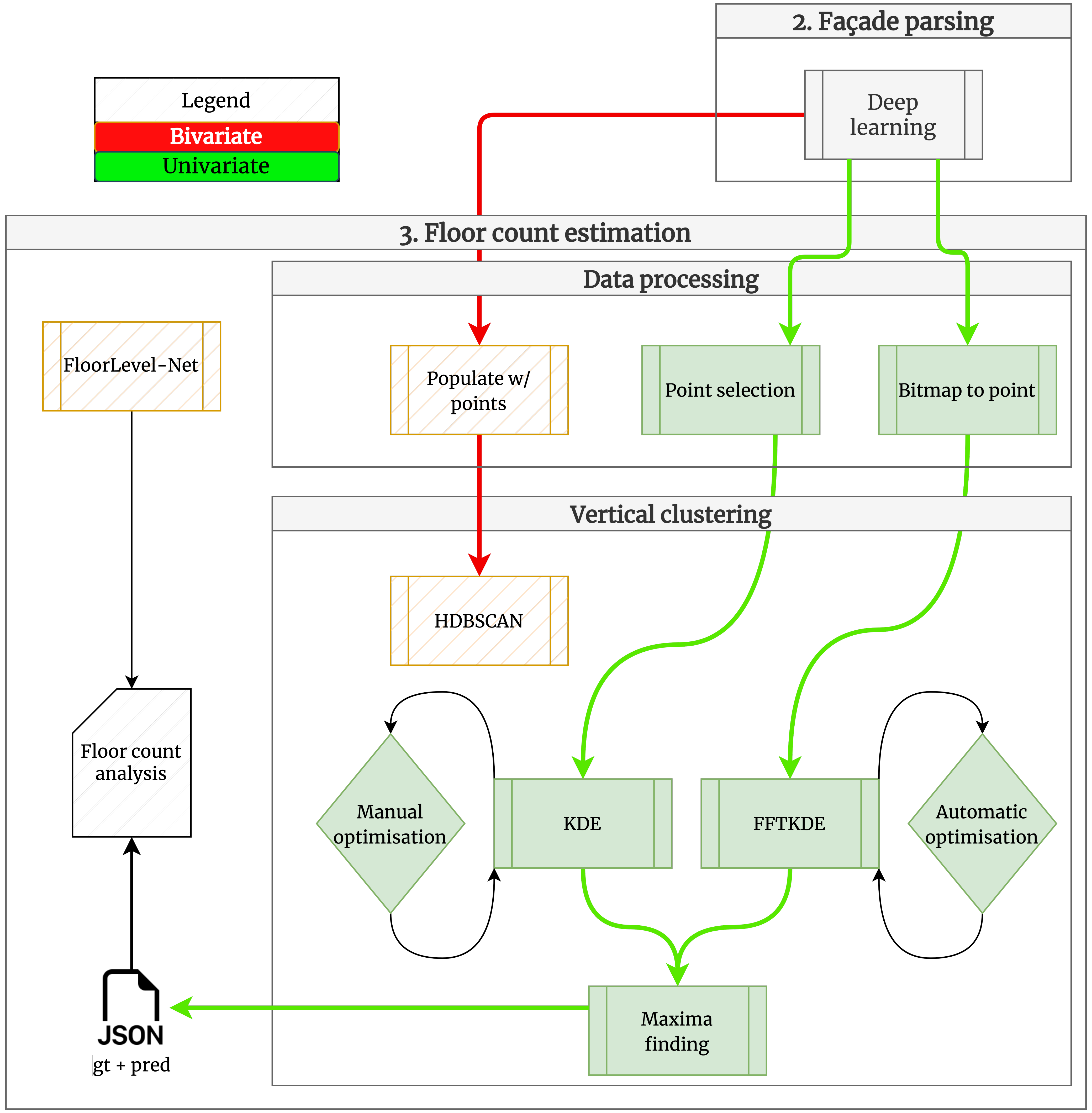
3. Floor count estimation

Data processing

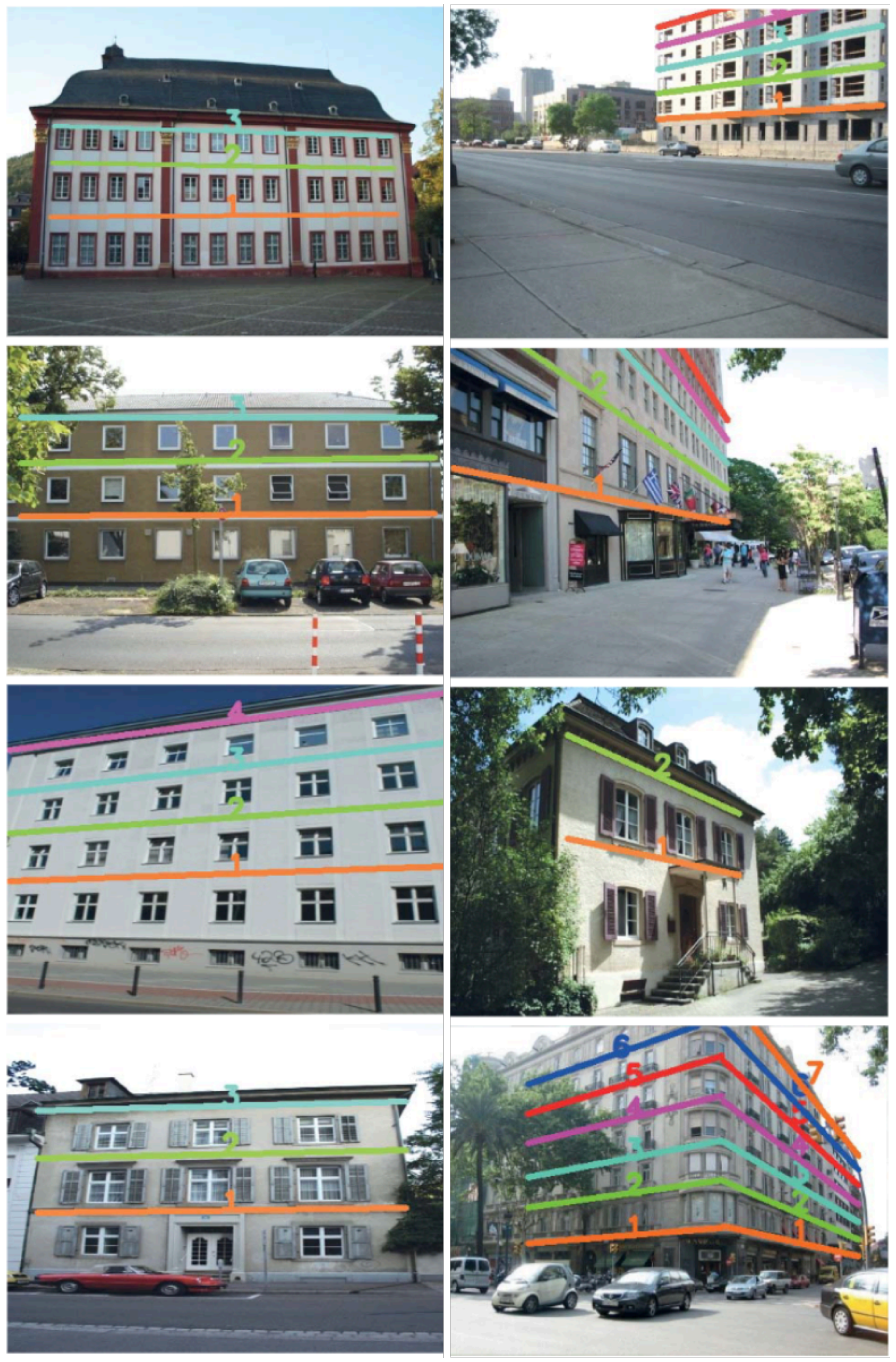
Legend

- 🟩 Successful / Good result
- 🟧 Successful / Poor result
- 🟥 Failure
- ▨ Discontinued / shelved

Floor count estimation

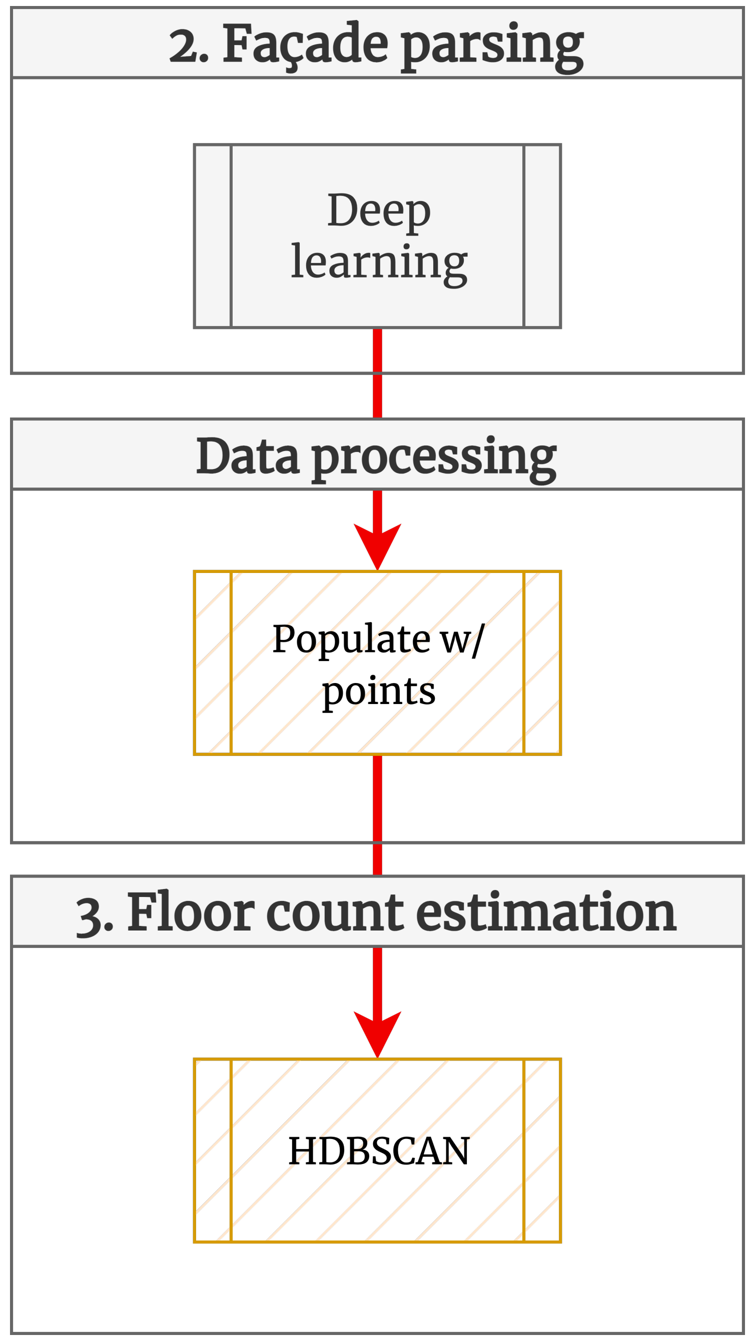


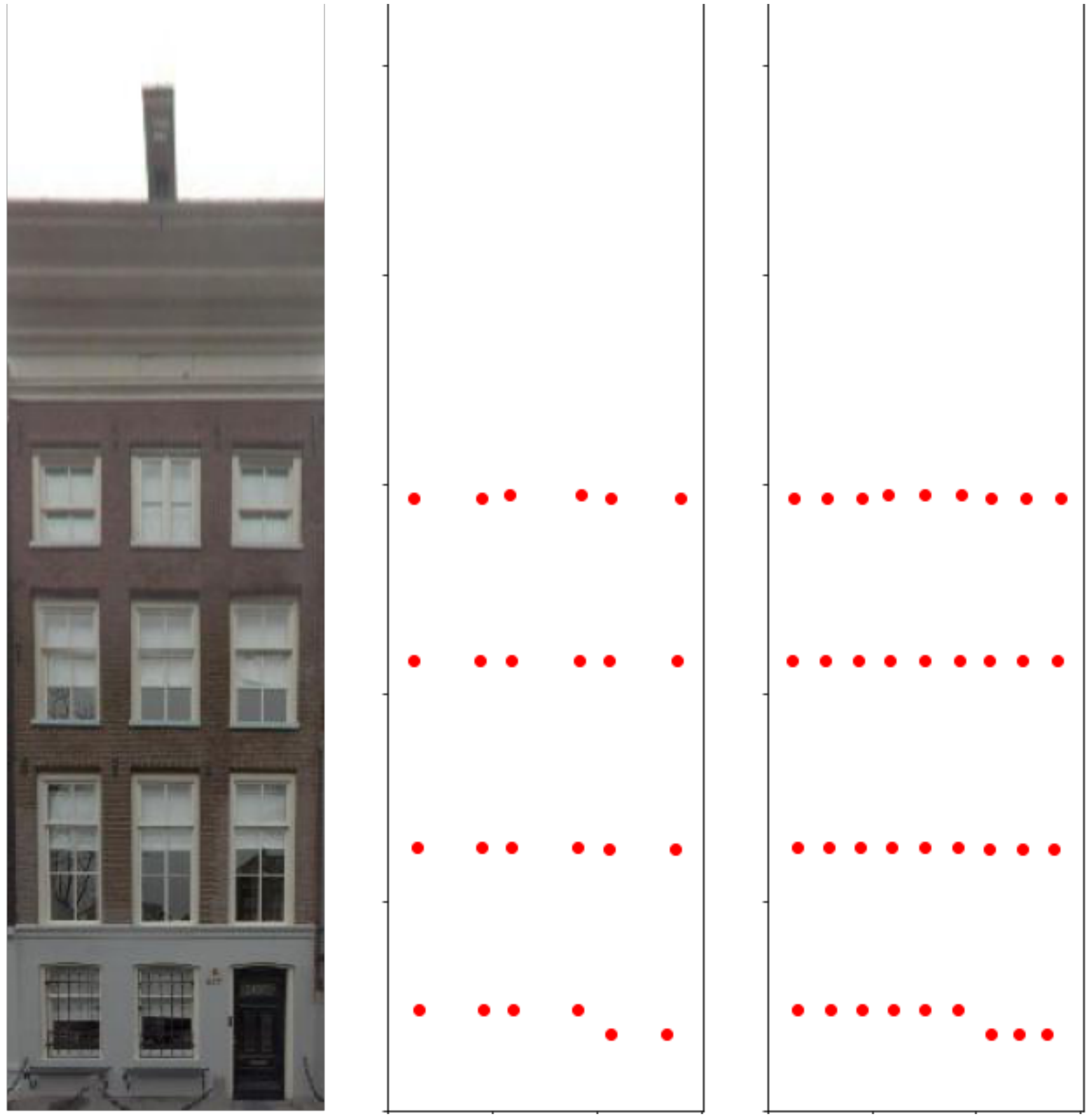
FloorLevel-Net



[12]

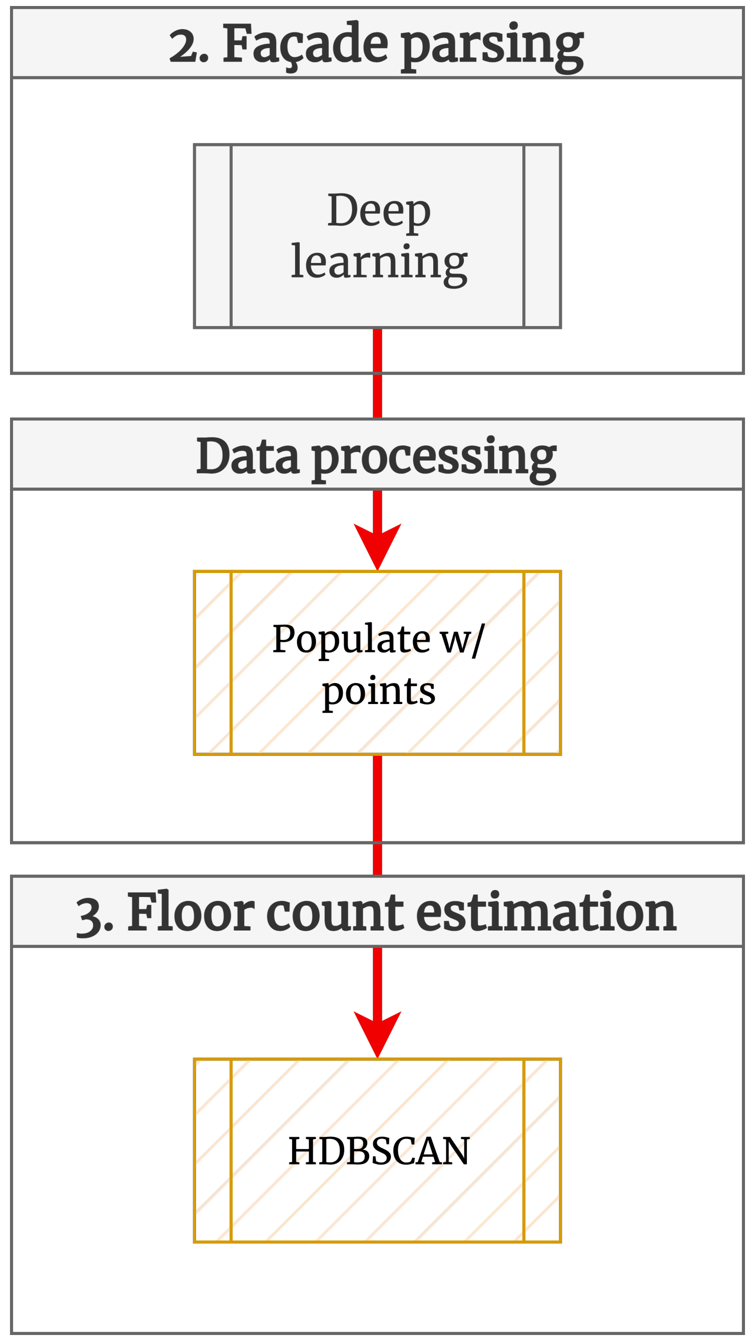
Façade parsing: Bivariate approach (x, y)





Façade parsing: Bivariate approach (x, y)

- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions



2. Façade parsing

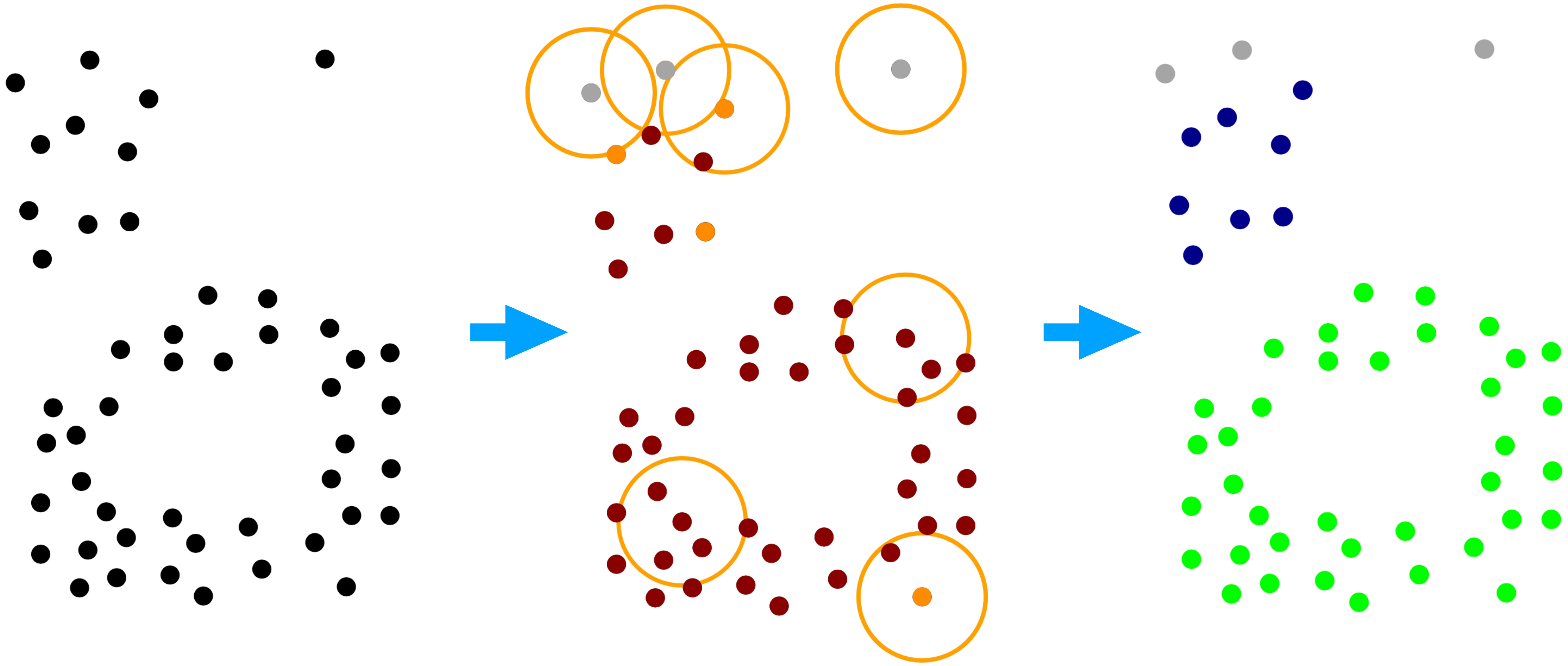
Deep learning

Data processing

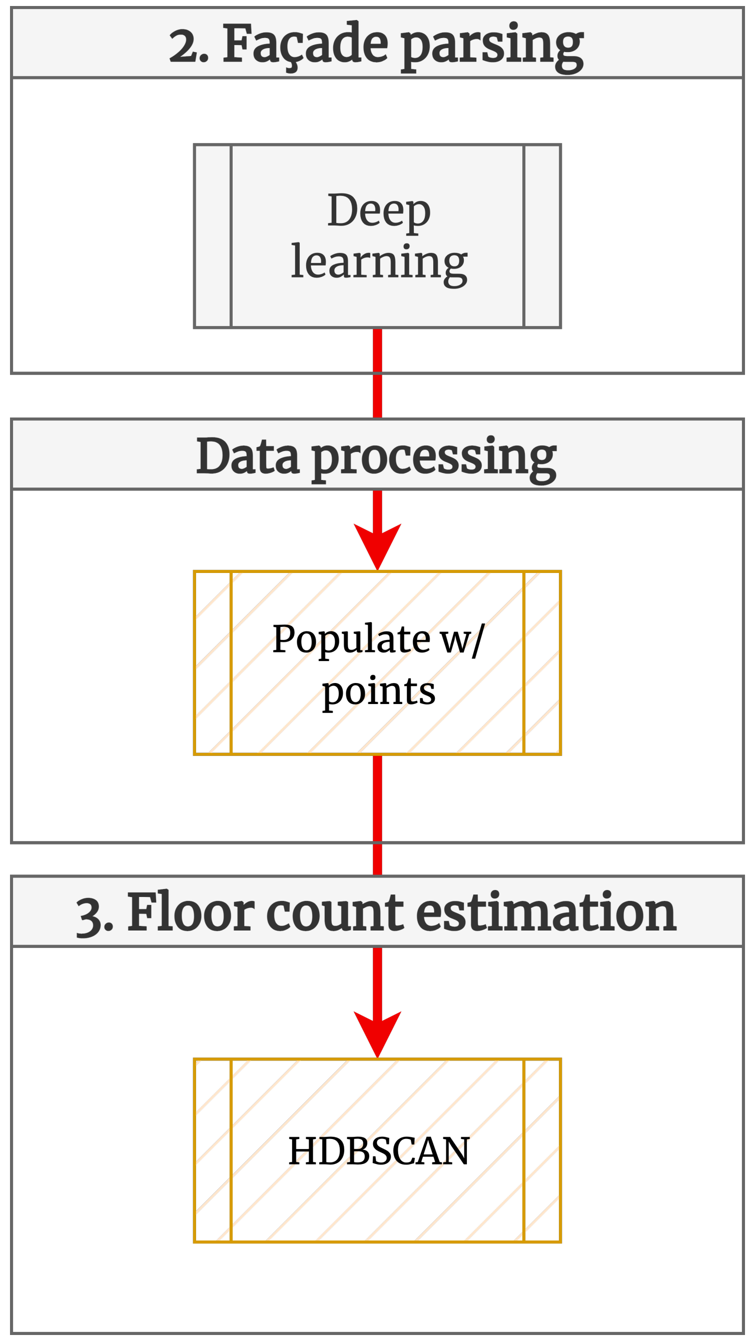
Populate w/ points

3. Floor count estimation

HDBSCAN



Façade parsing: Bivariate approach (x, y)



2. Façade parsing

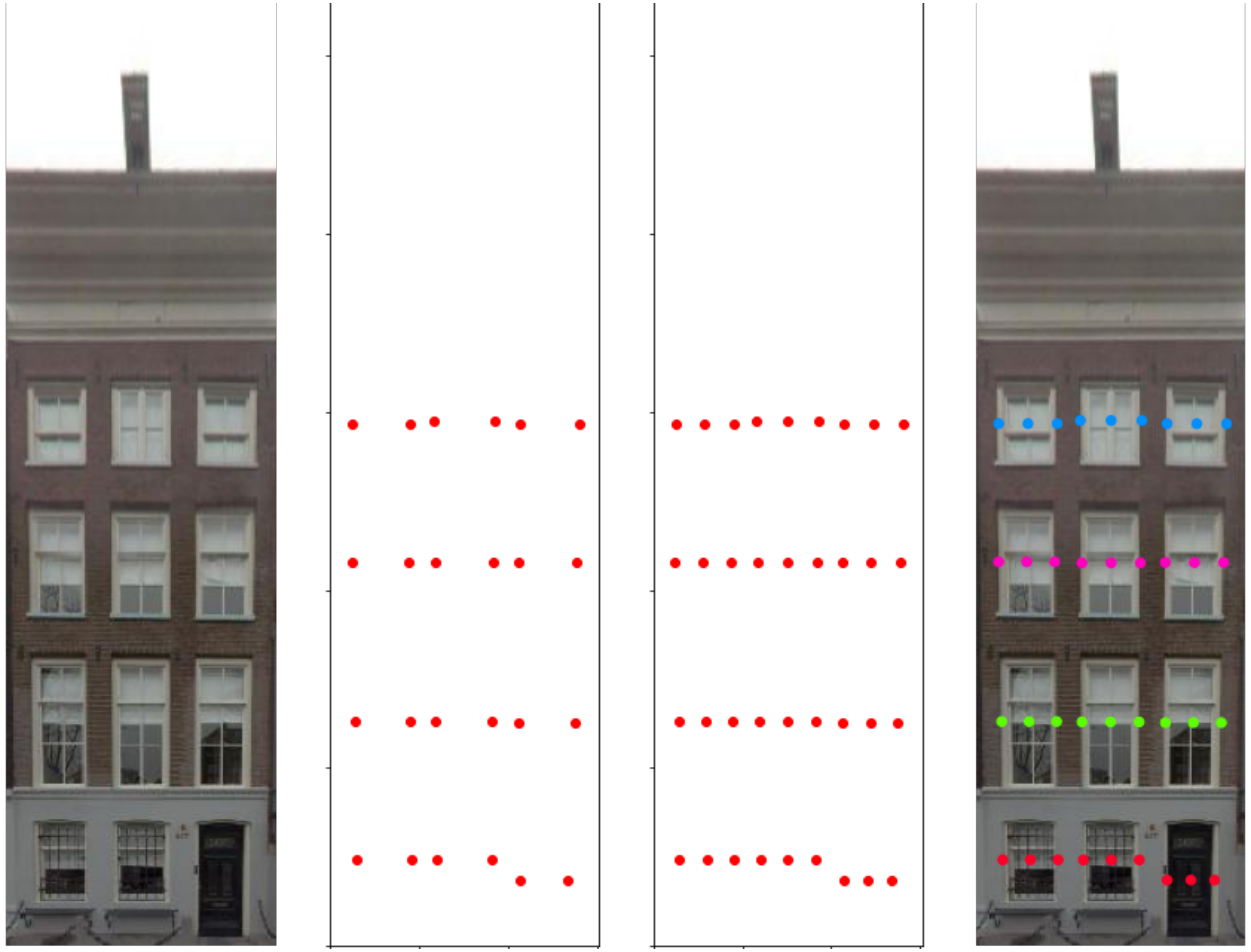
Deep learning

Data processing

Populate w/ points

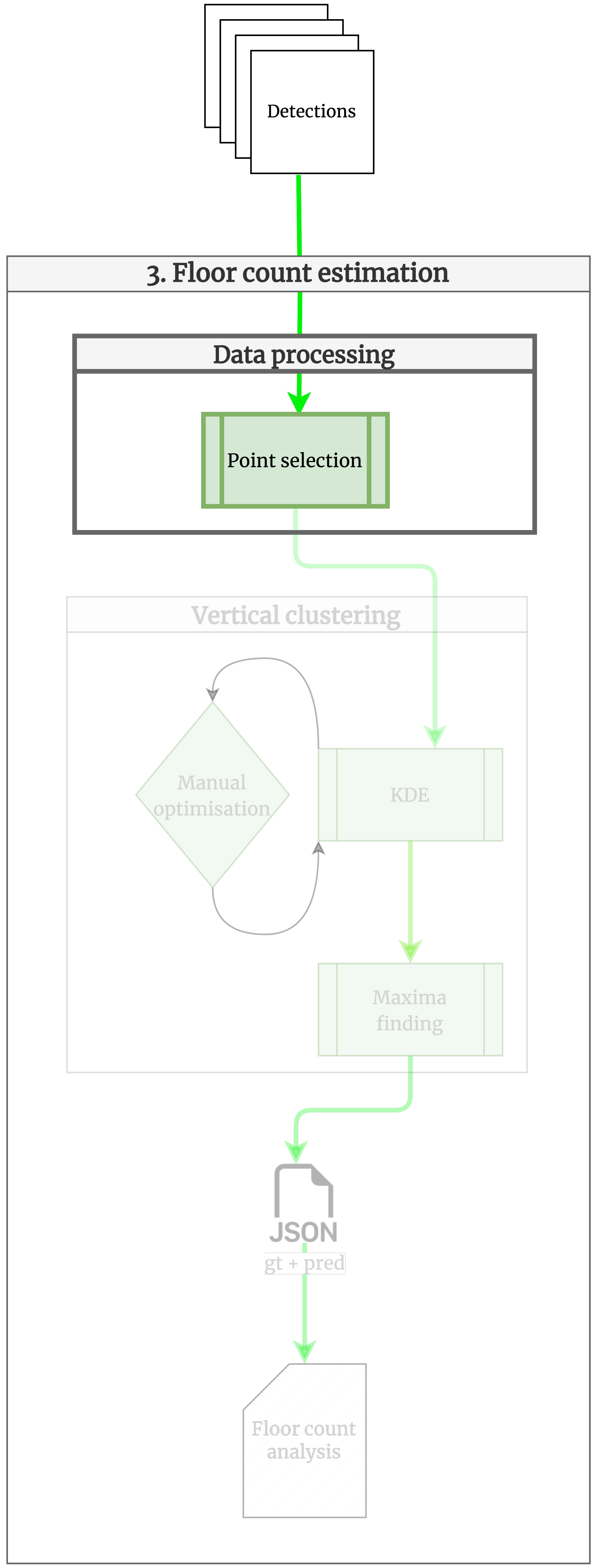
3. Floor count estimation

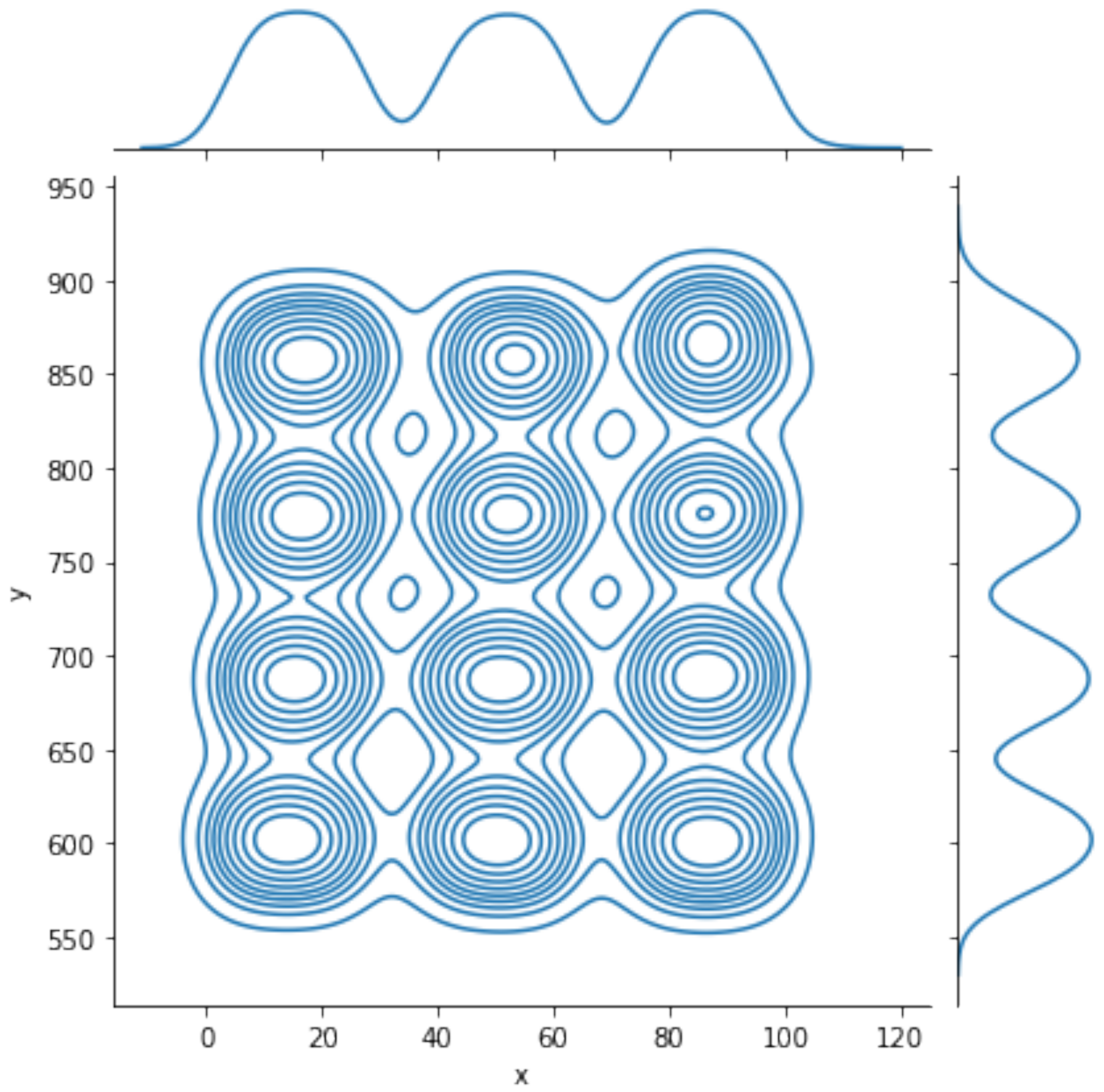
HDBSCAN



Façade parsing: Univariate approach (y)

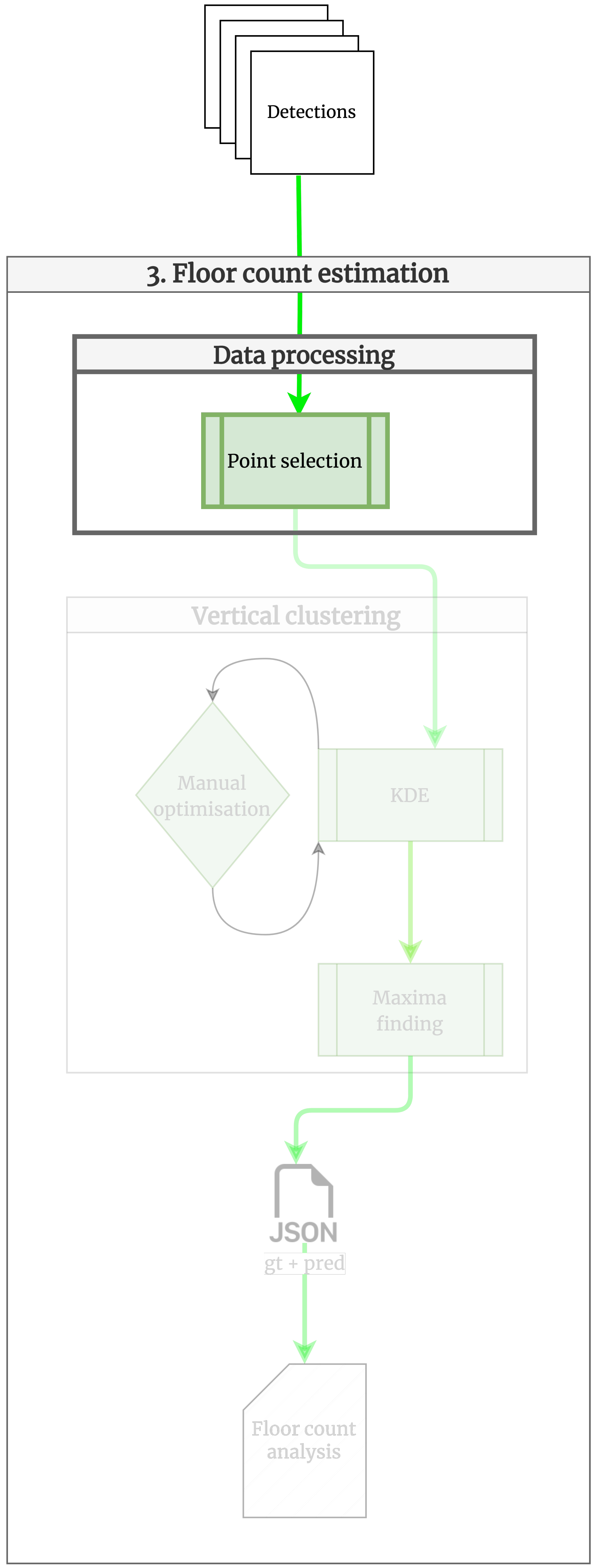
- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions





Façade parsing: Univariate approach (y)

- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions



Detections

3. Floor count estimation

Data processing

Point selection

Vertical clustering

Manual optimisation

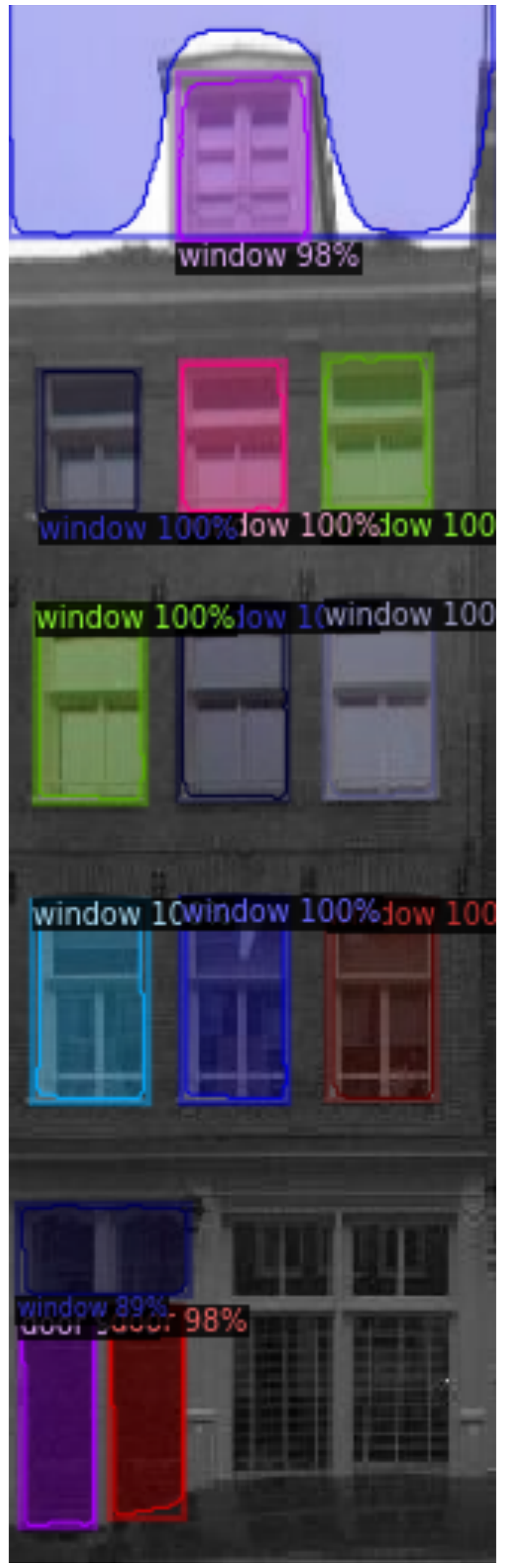
KDE

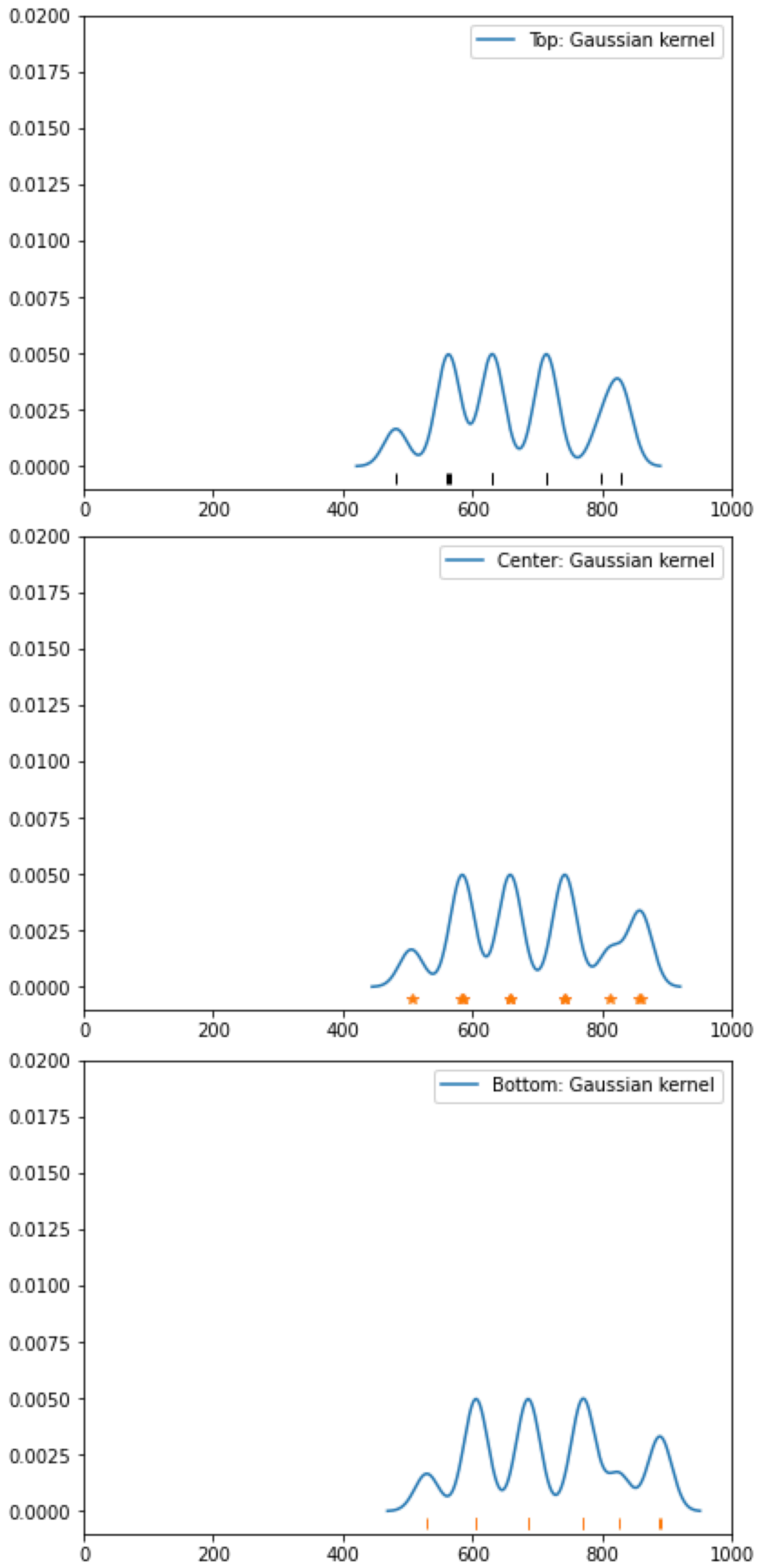
Maxima finding

JSON

gt + pred

Floor count analysis

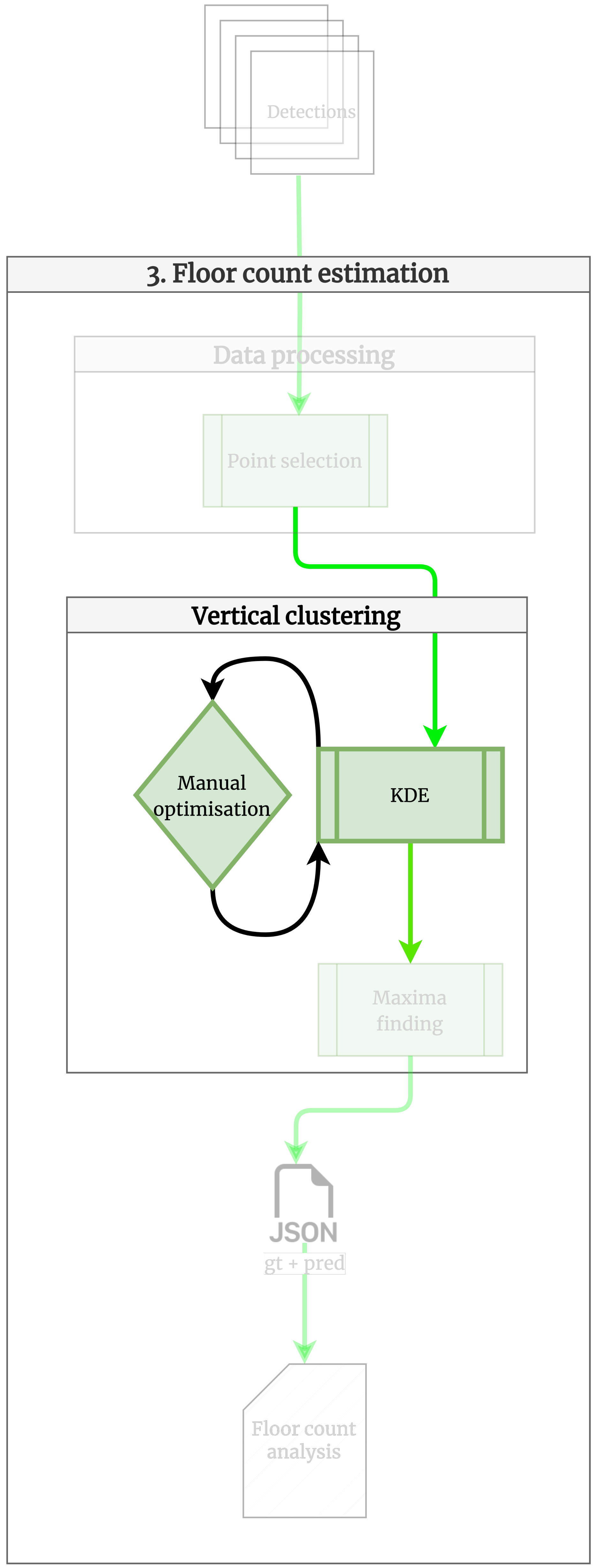




Façade parsing: Univariate approach (y)

- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

TUDelft



Detections

3. Floor count estimation

Data processing

Point selection

Vertical clustering

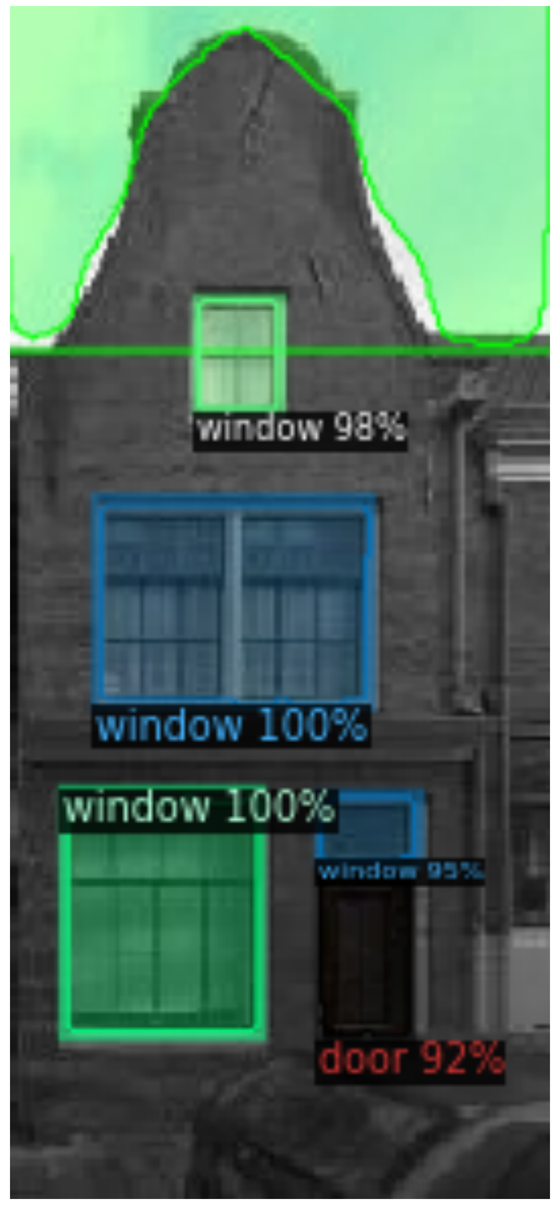
Manual optimisation

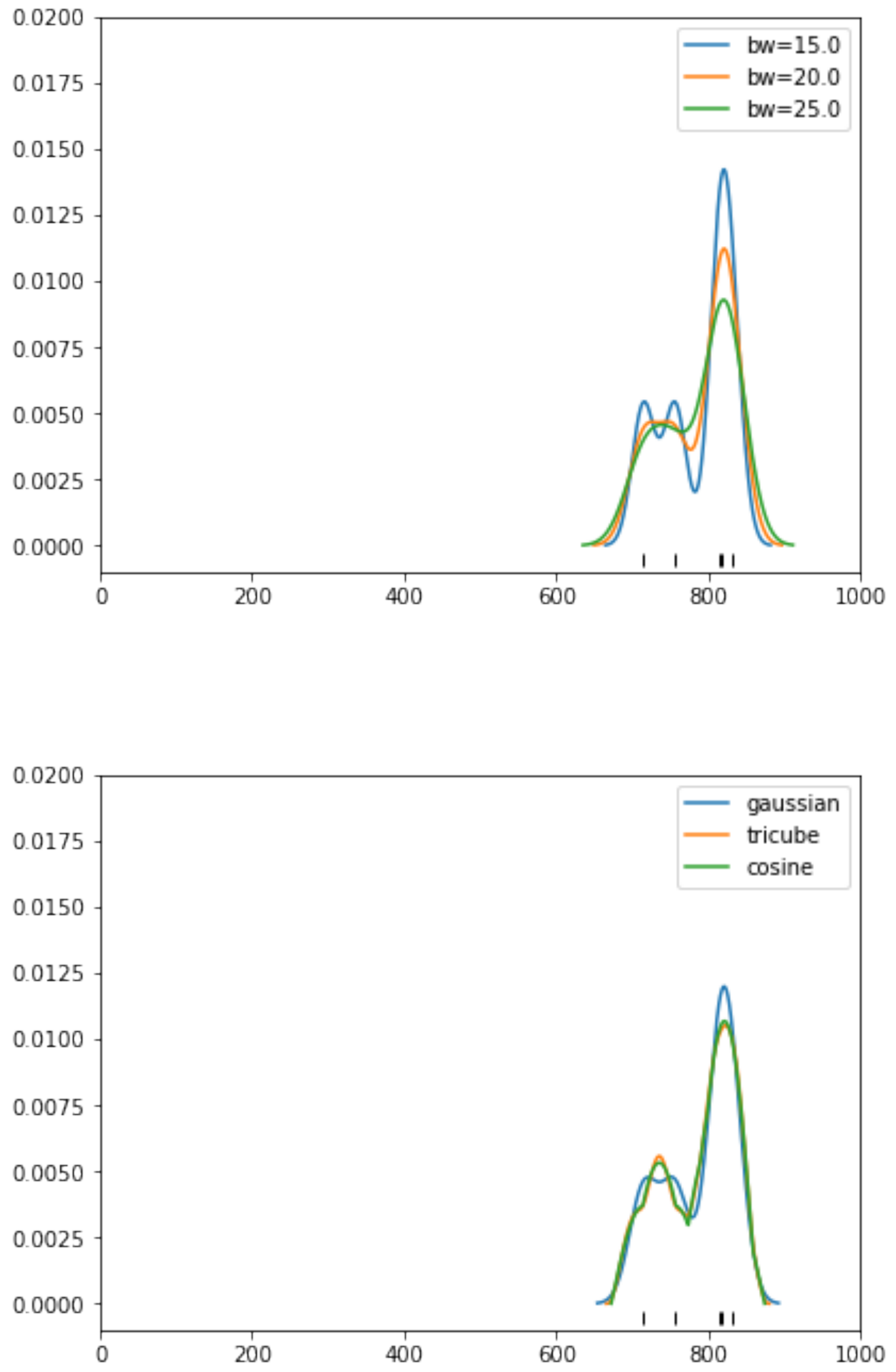
KDE

Maxima finding

JSON gt + pred

Floor count analysis

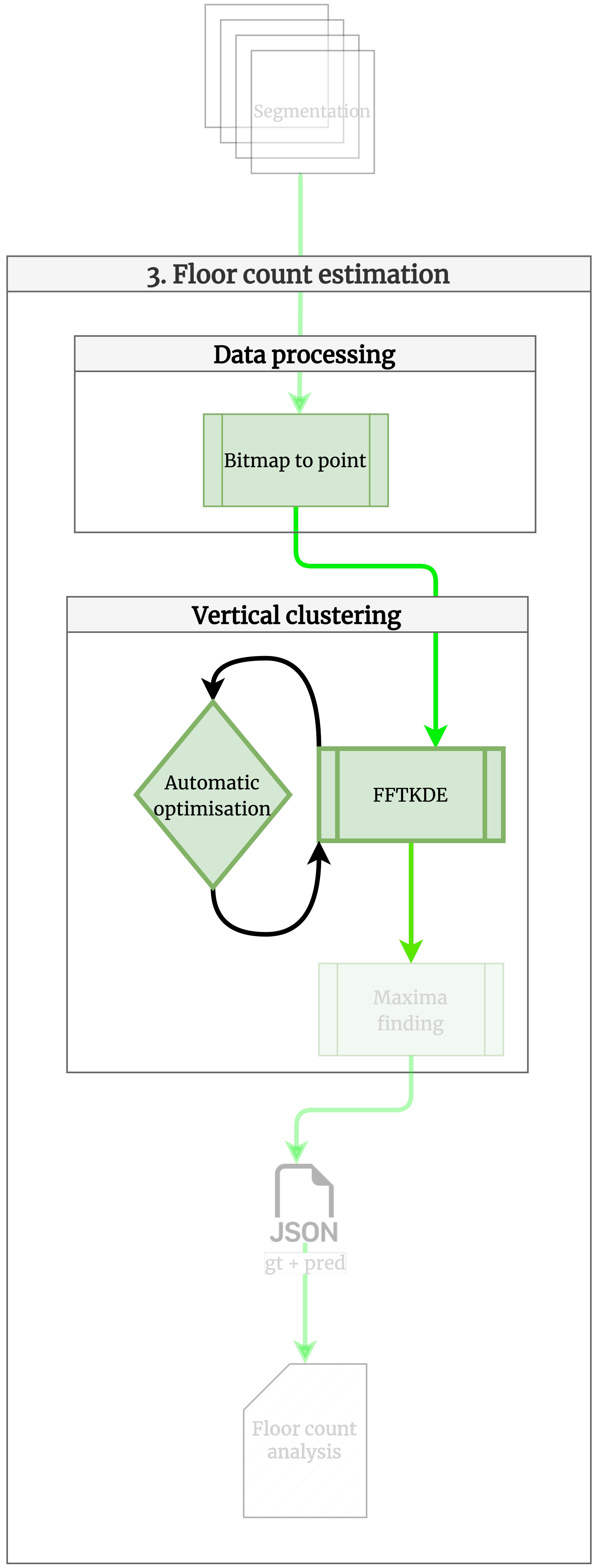


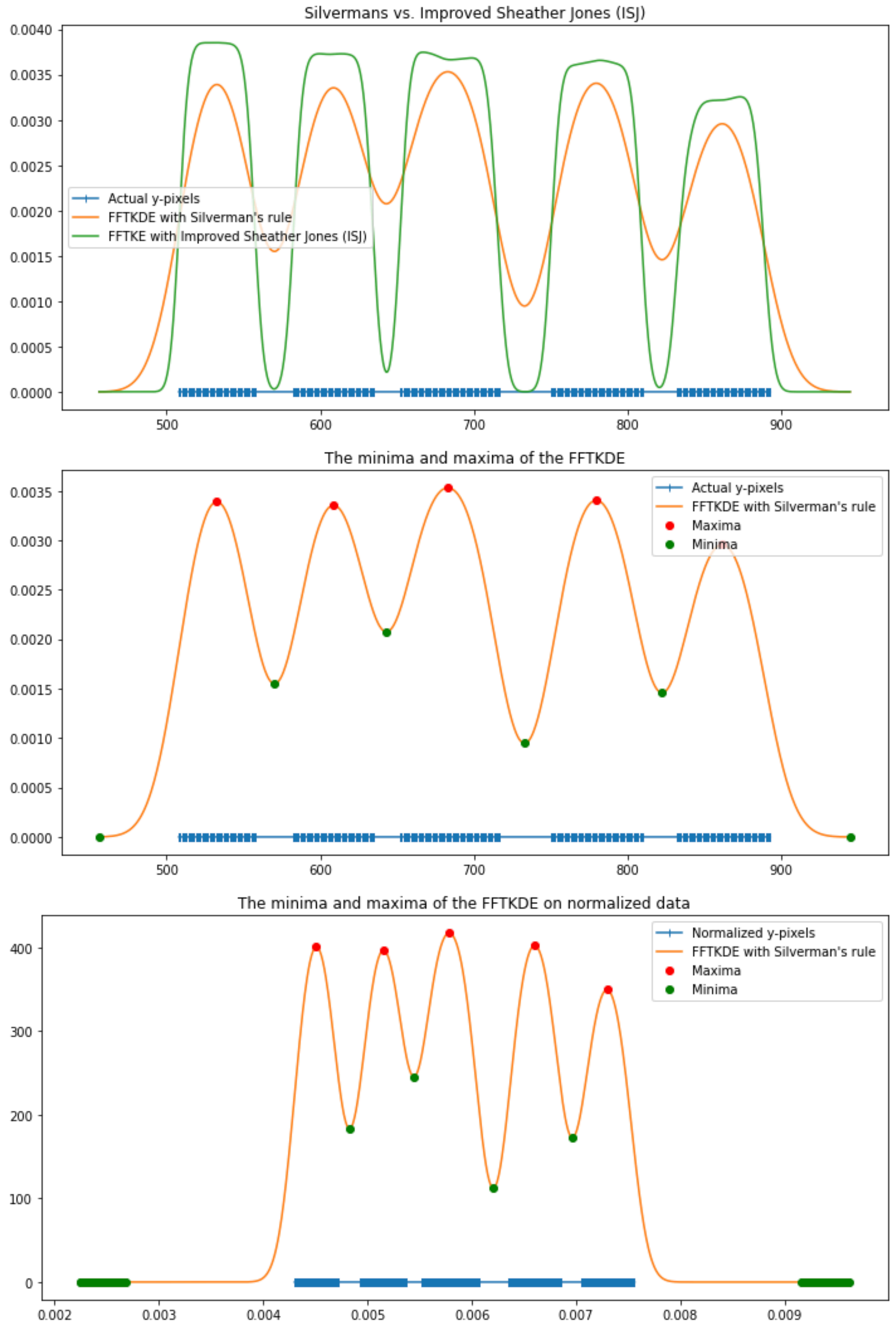


Façade parsing: Univariate approach (y)

- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

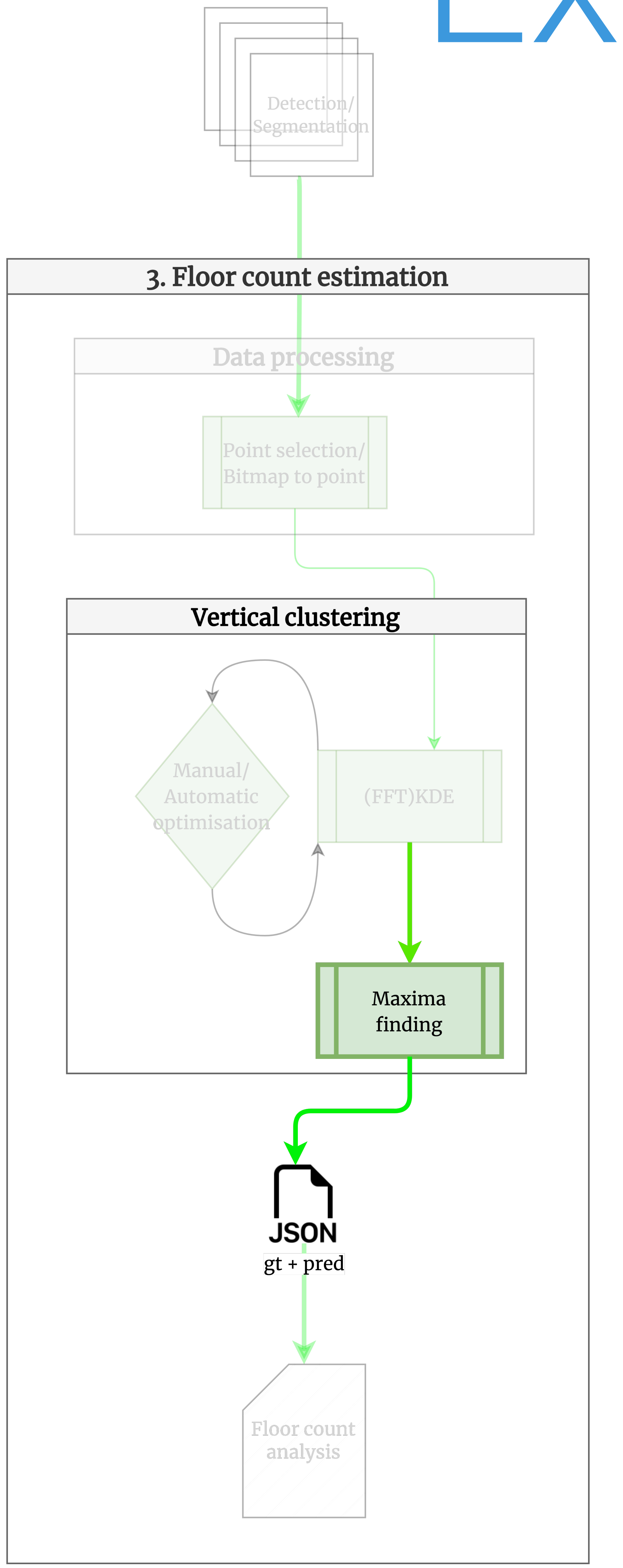
TU Delft





- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

Extracting floor count



Detection/ Segmentation

3. Floor count estimation

Data processing

Point selection/ Bitmap to point

Vertical clustering

Manual/ Automatic optimisation

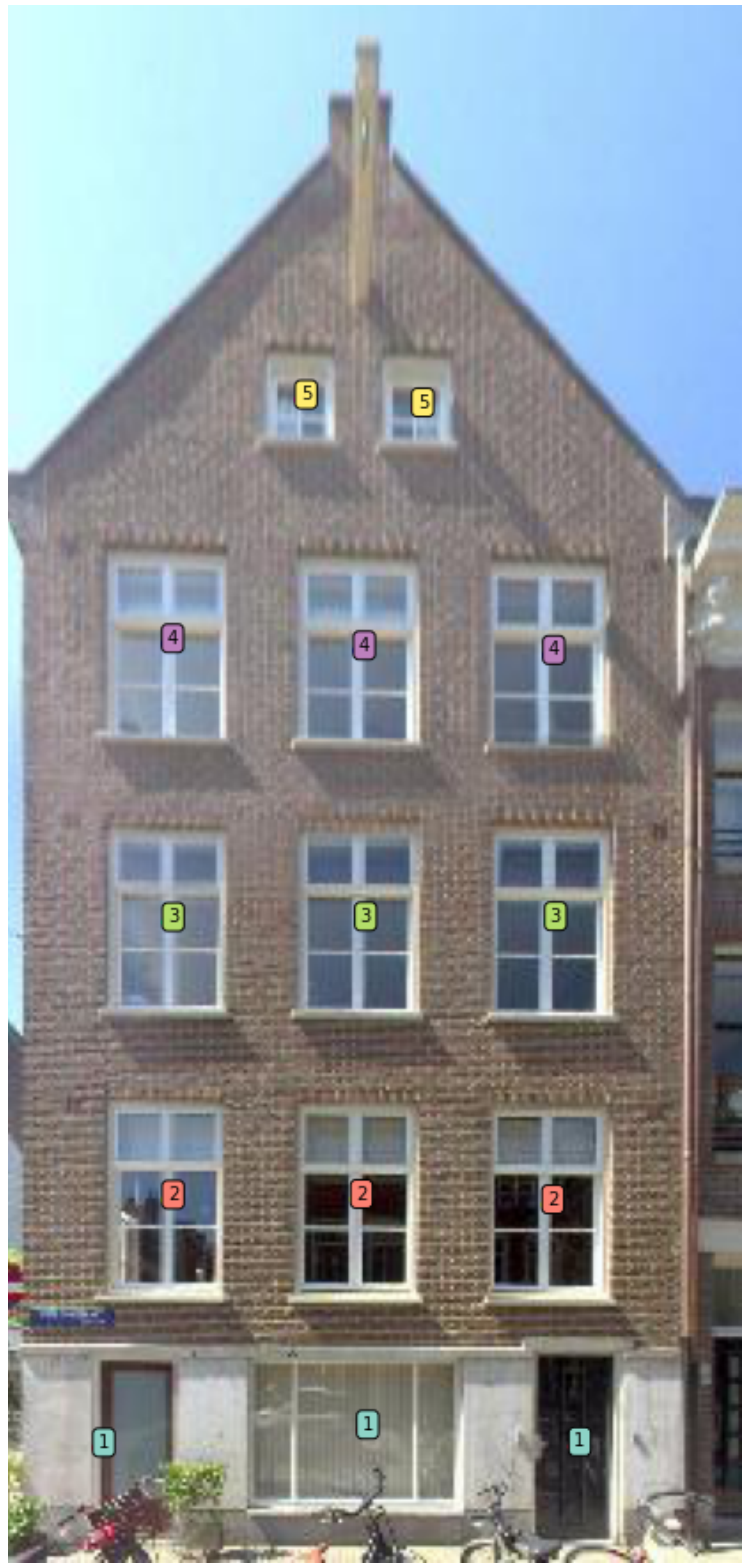
(FFT)KDE

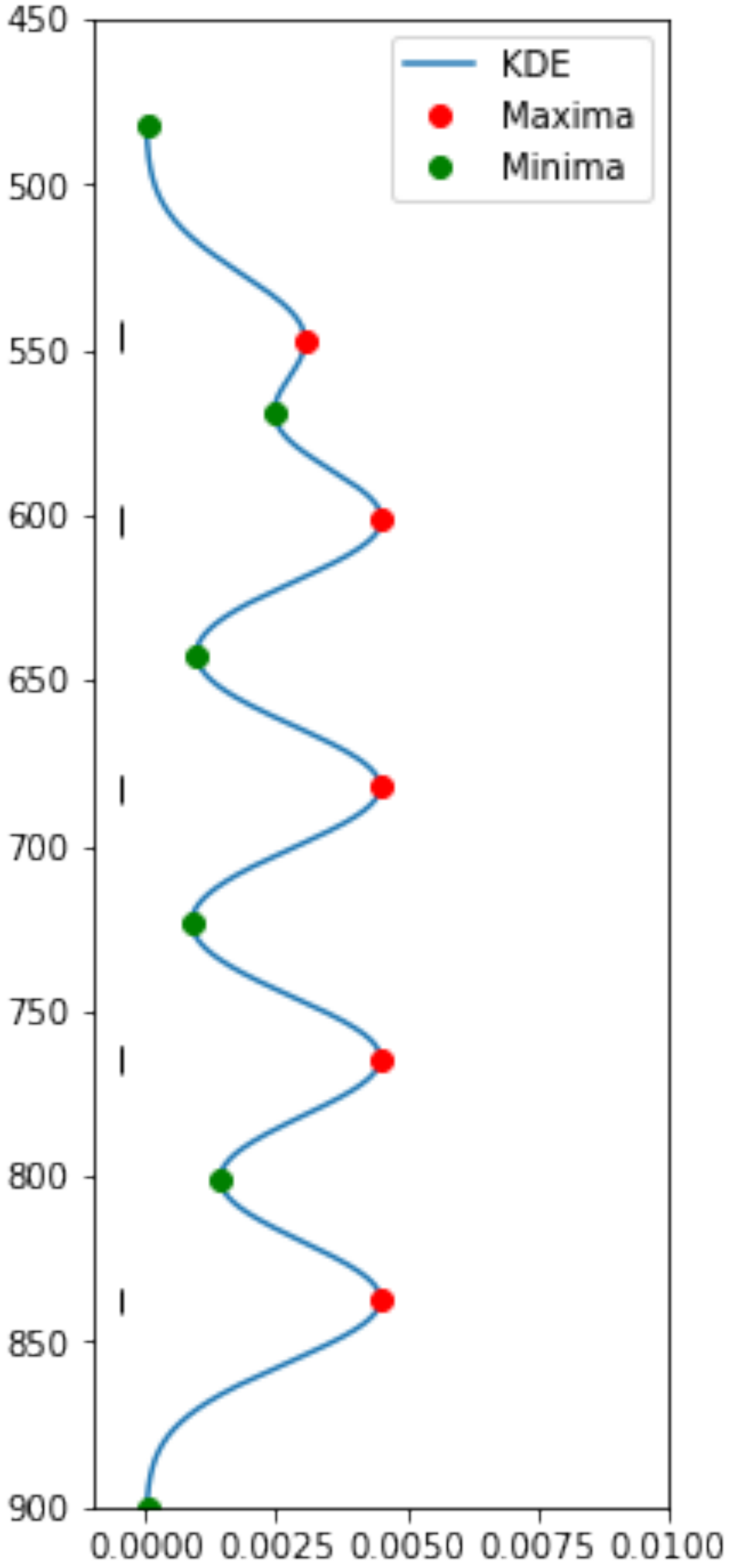
Maxima finding

JSON

gt + pred

Floor count analysis

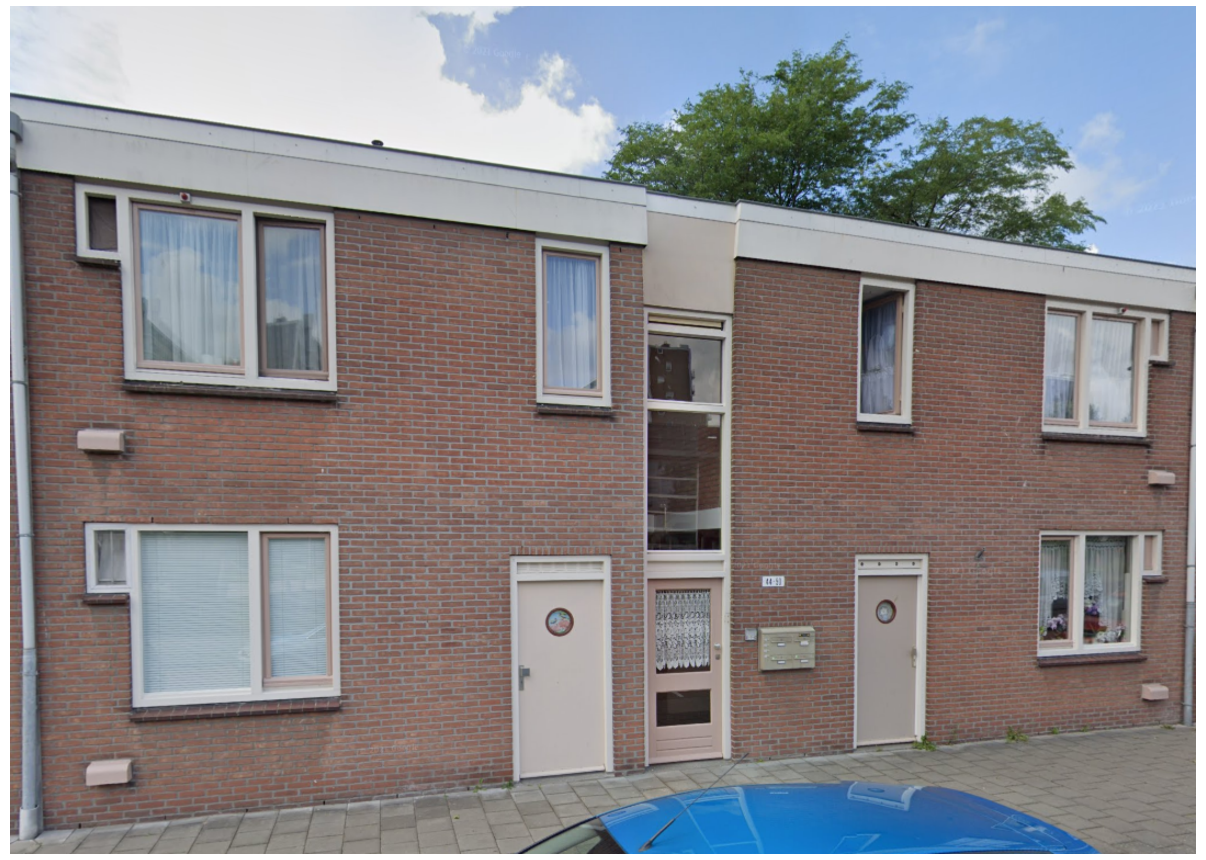


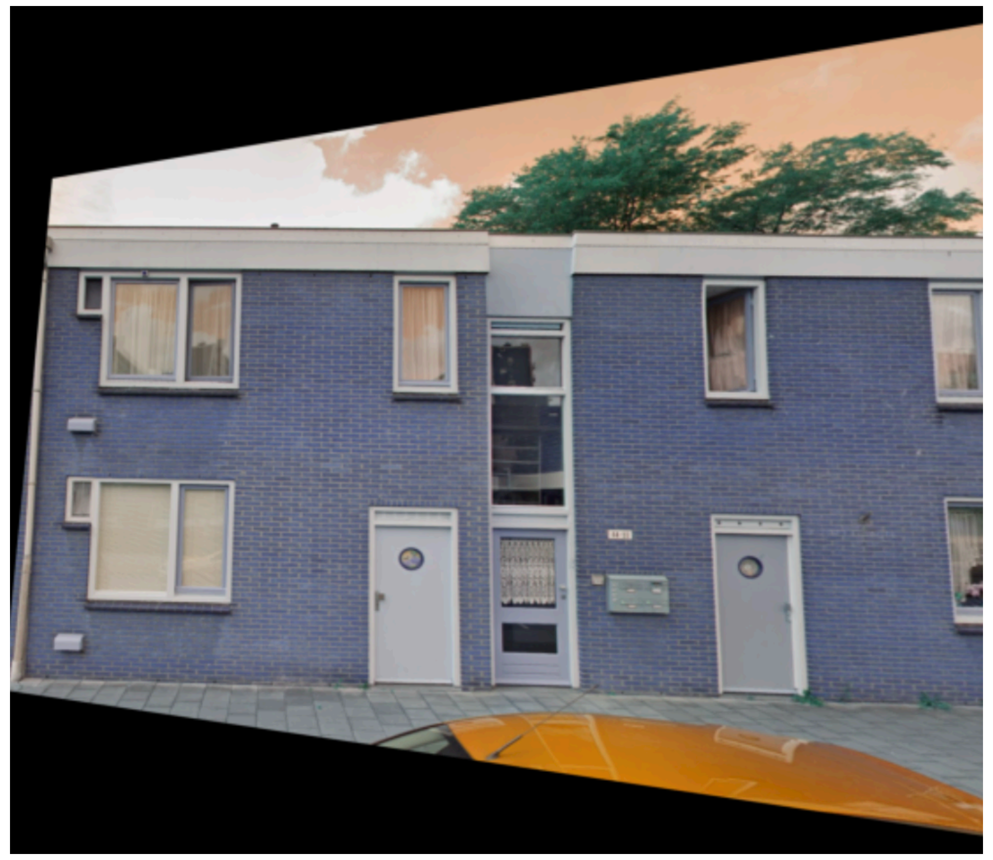


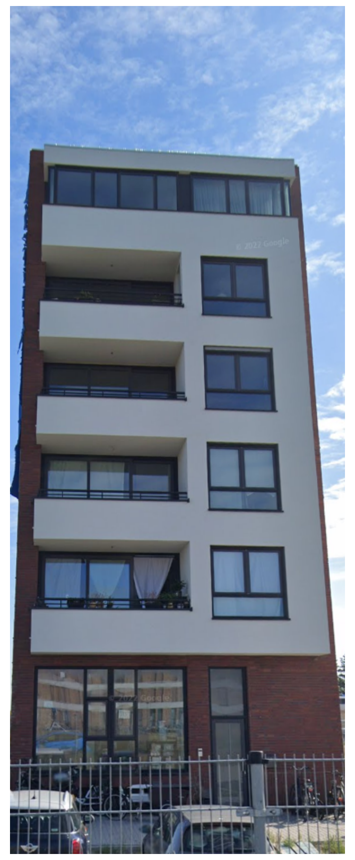
Results & Analysis

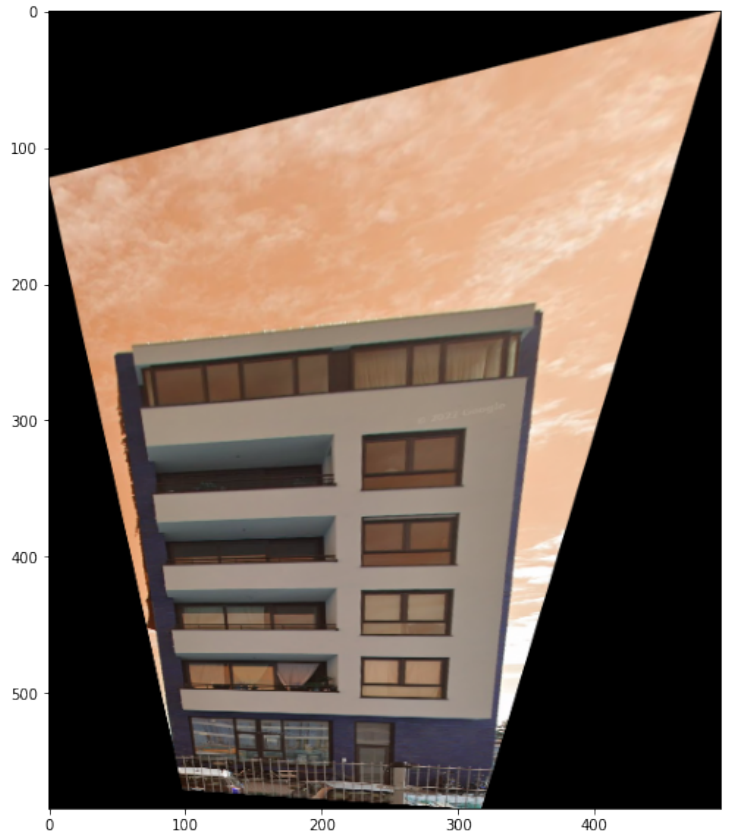
- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

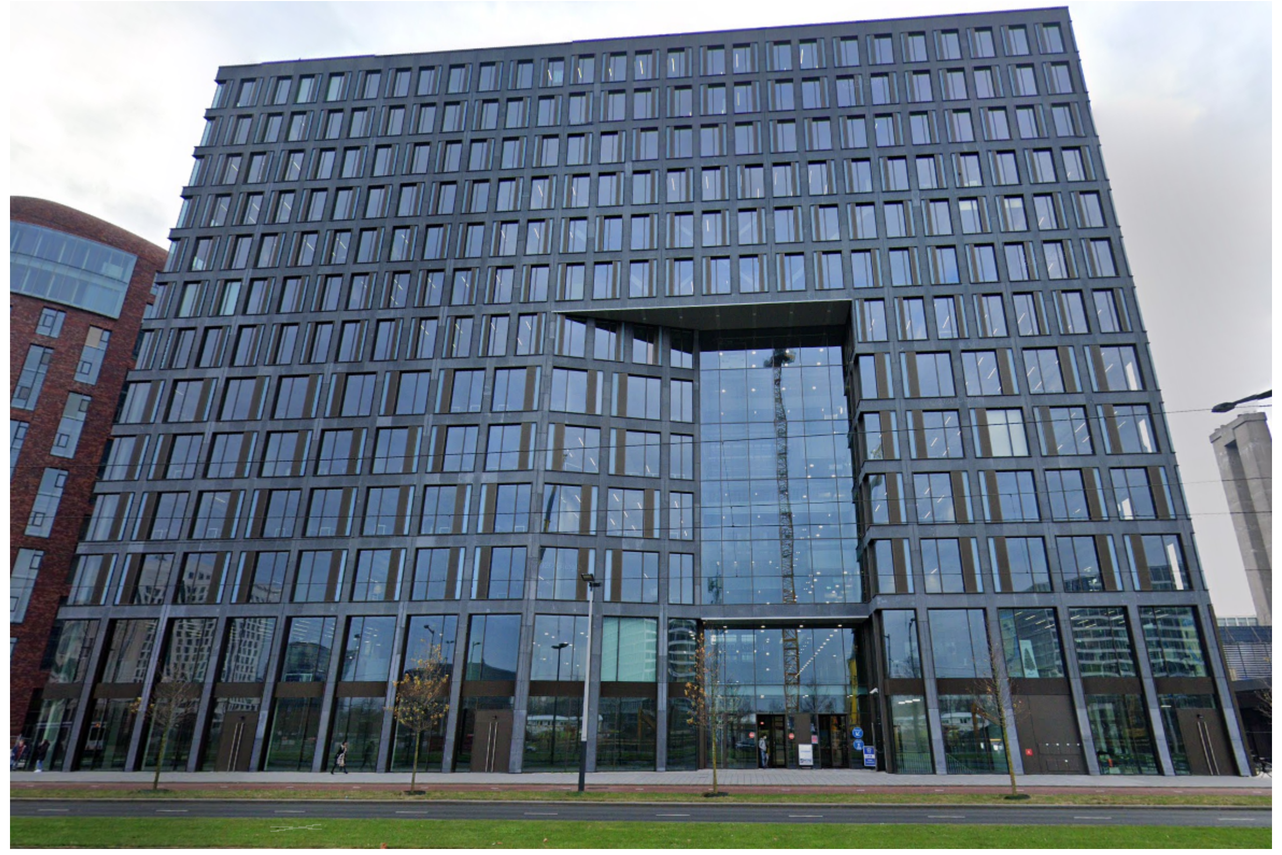
Image rectification: VP estimation

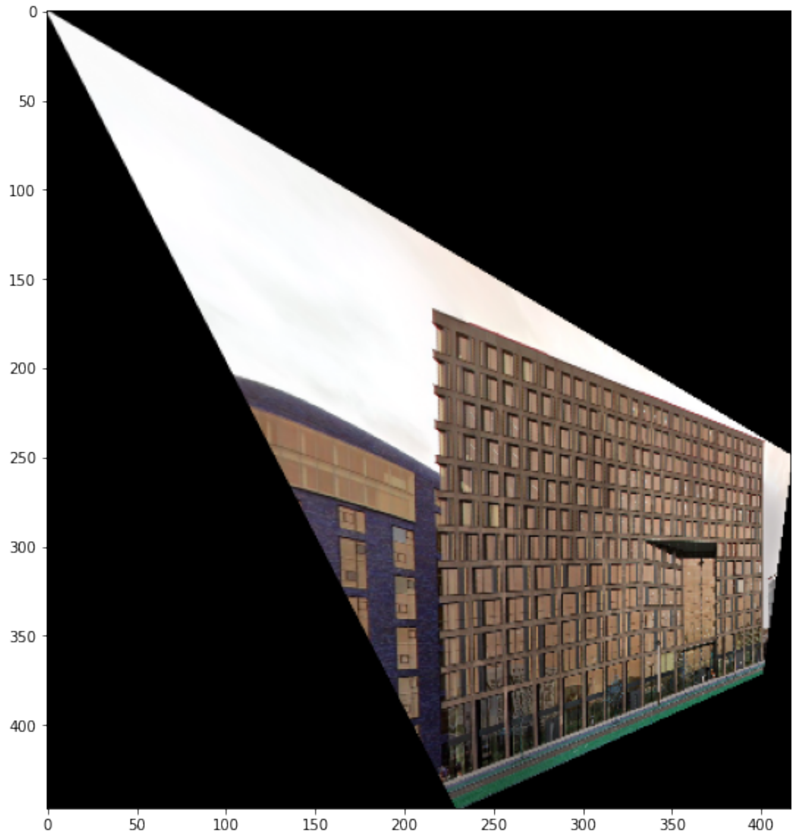


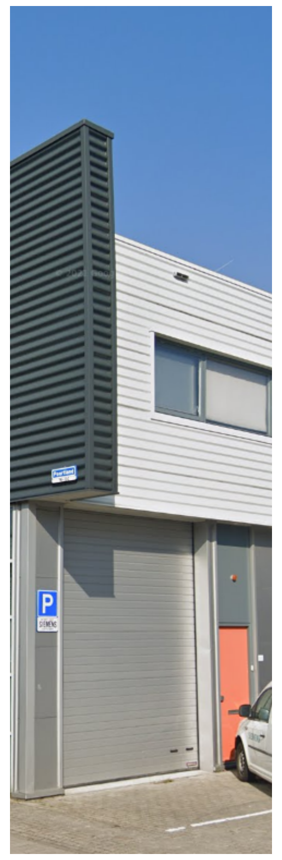


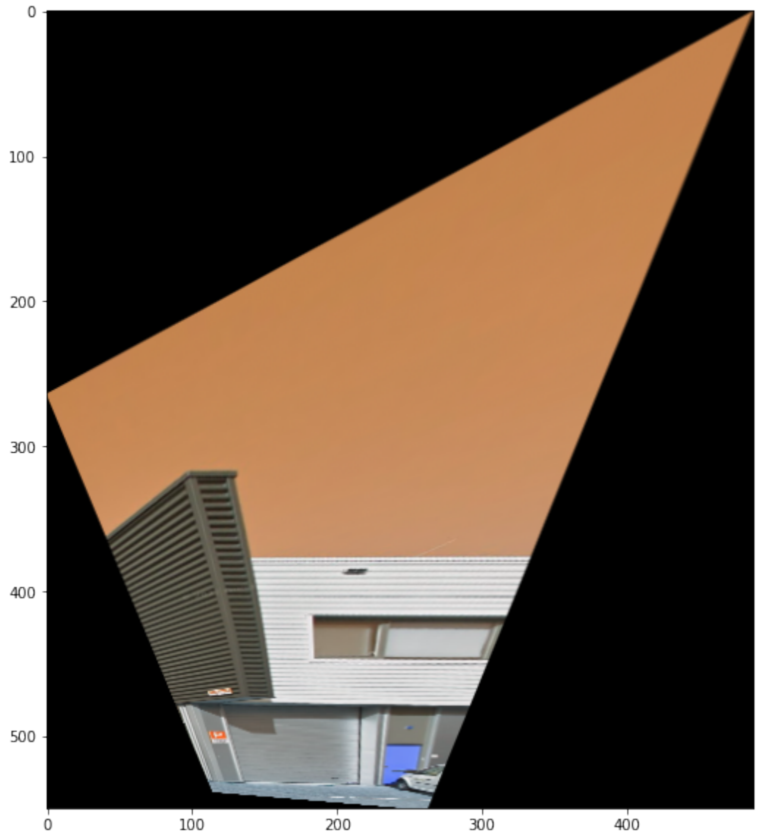






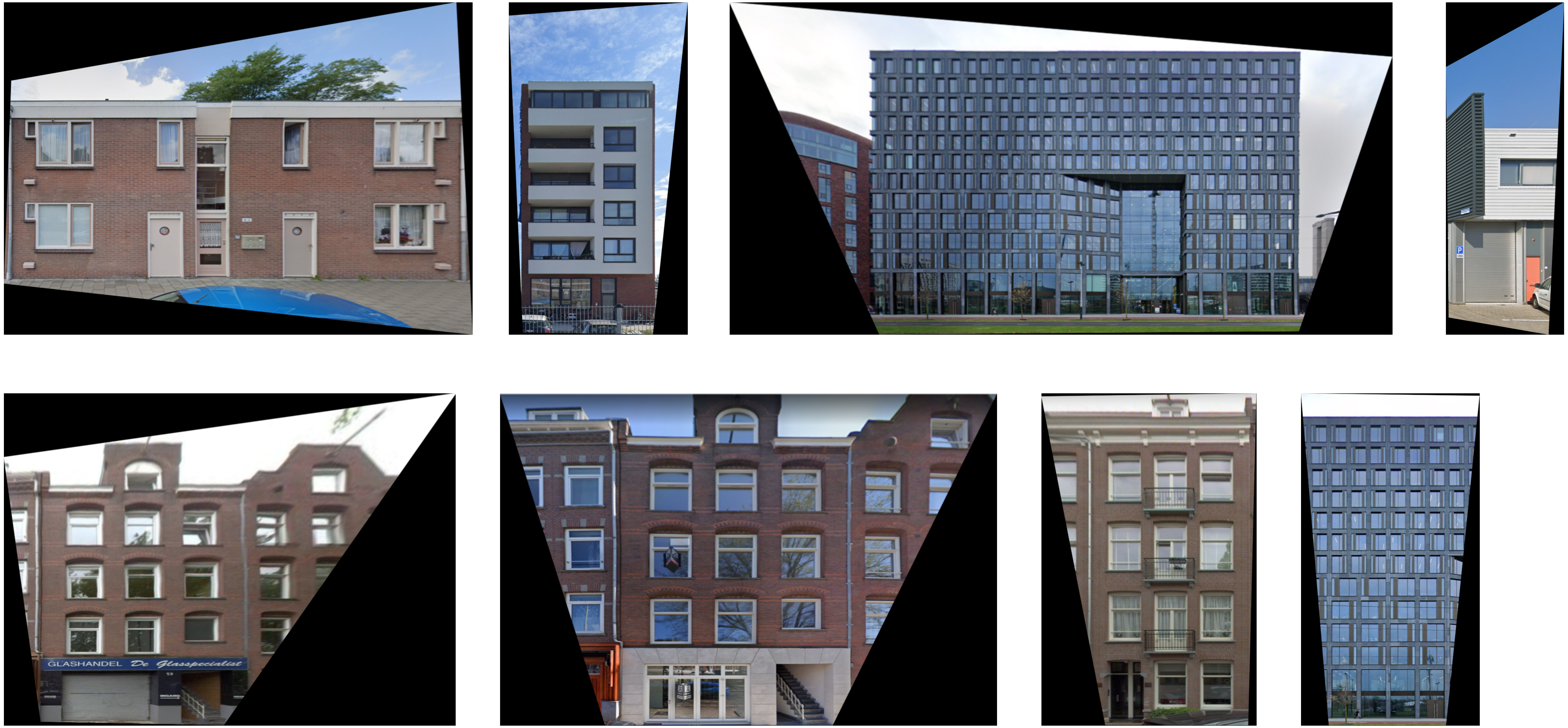






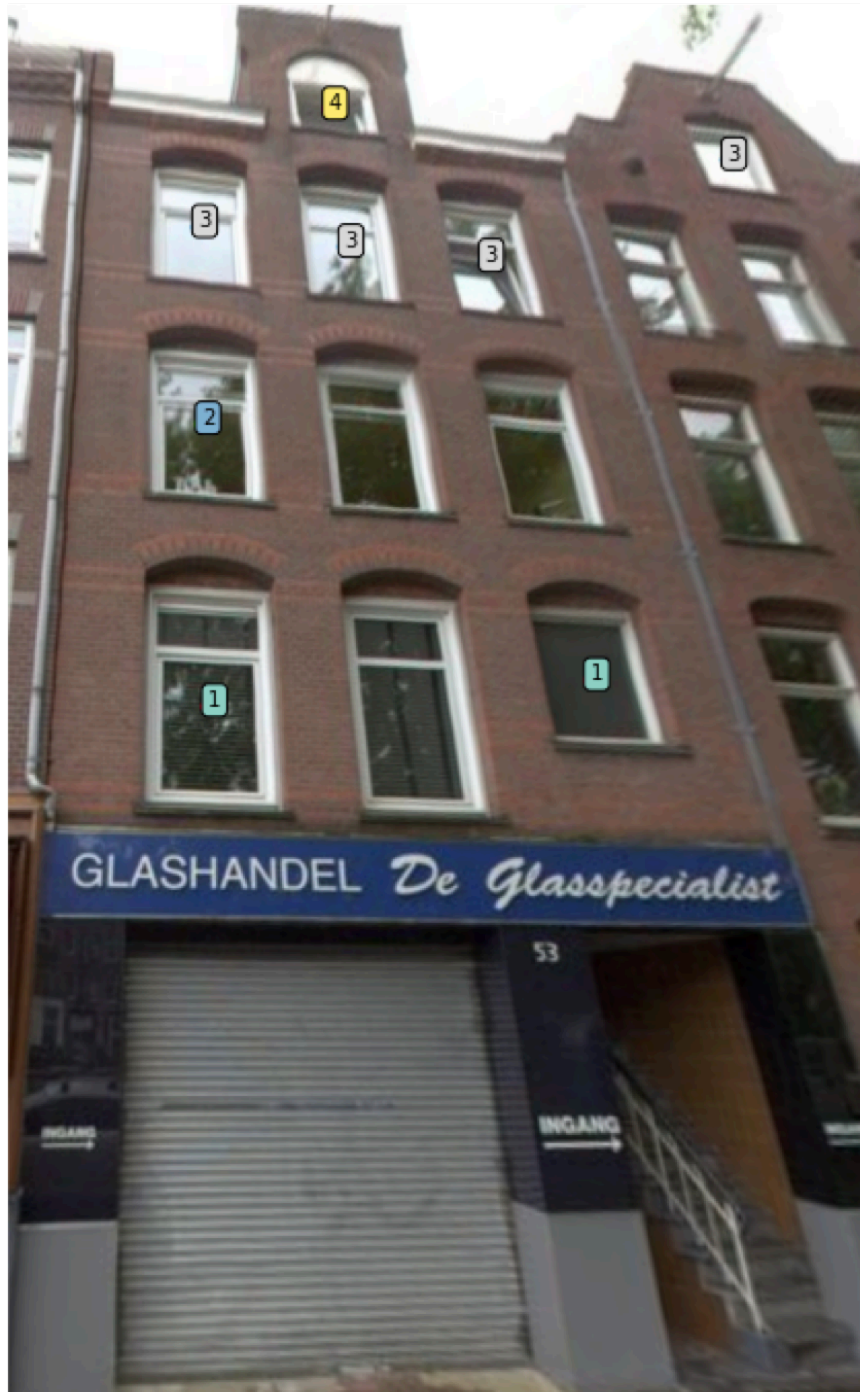
- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

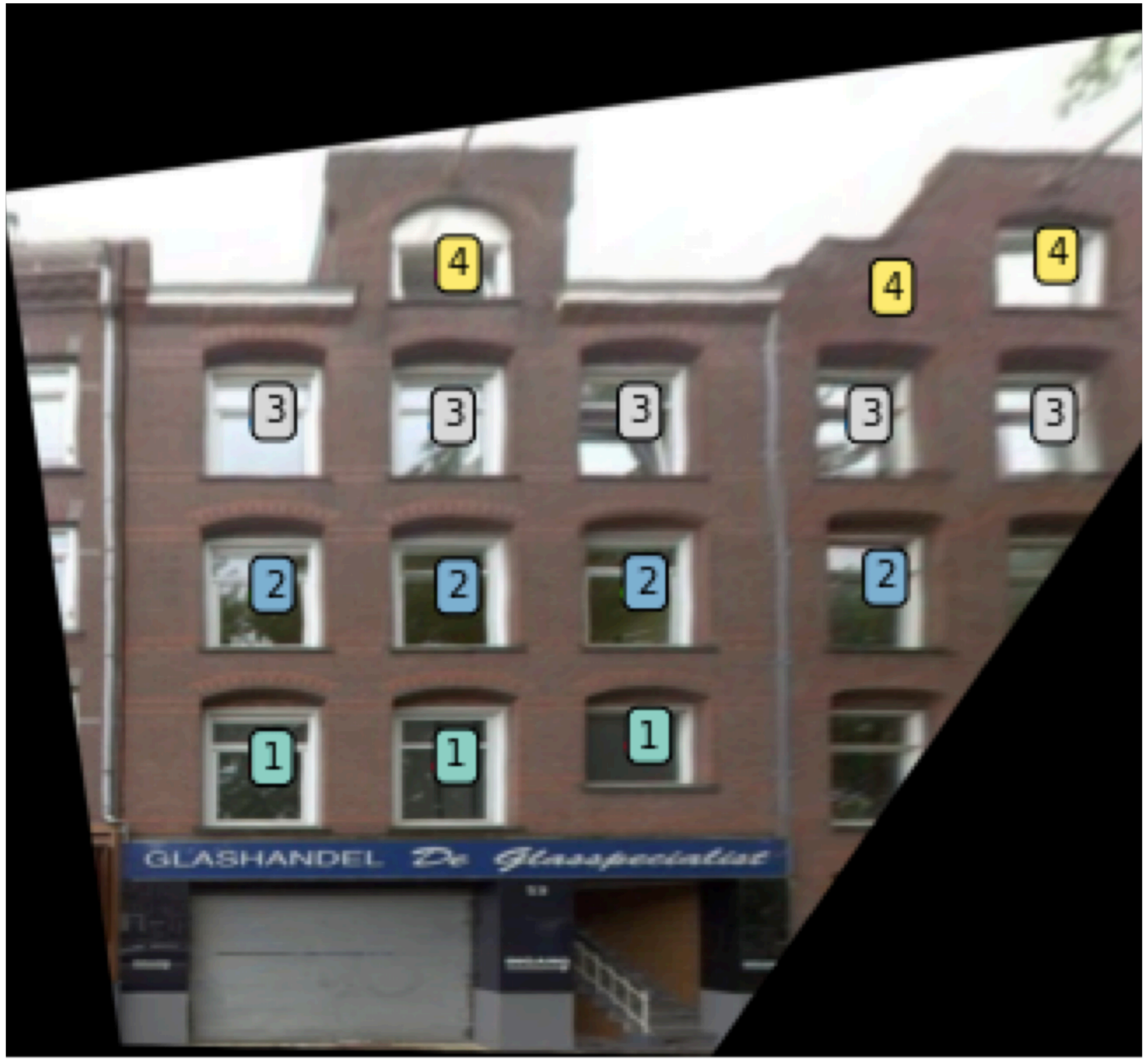
Image rectification: direct H transform



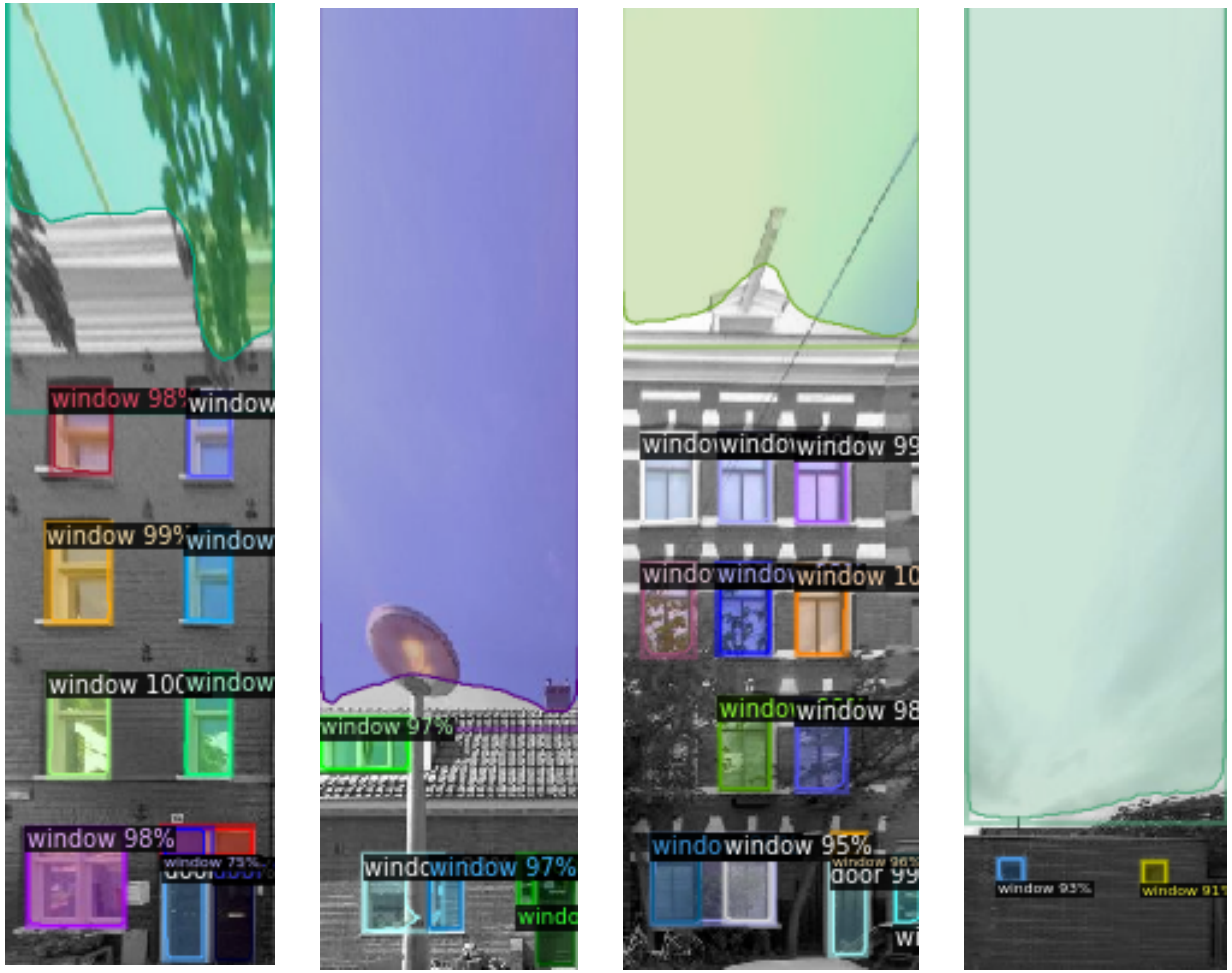
- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

Image rectification: direct H transform





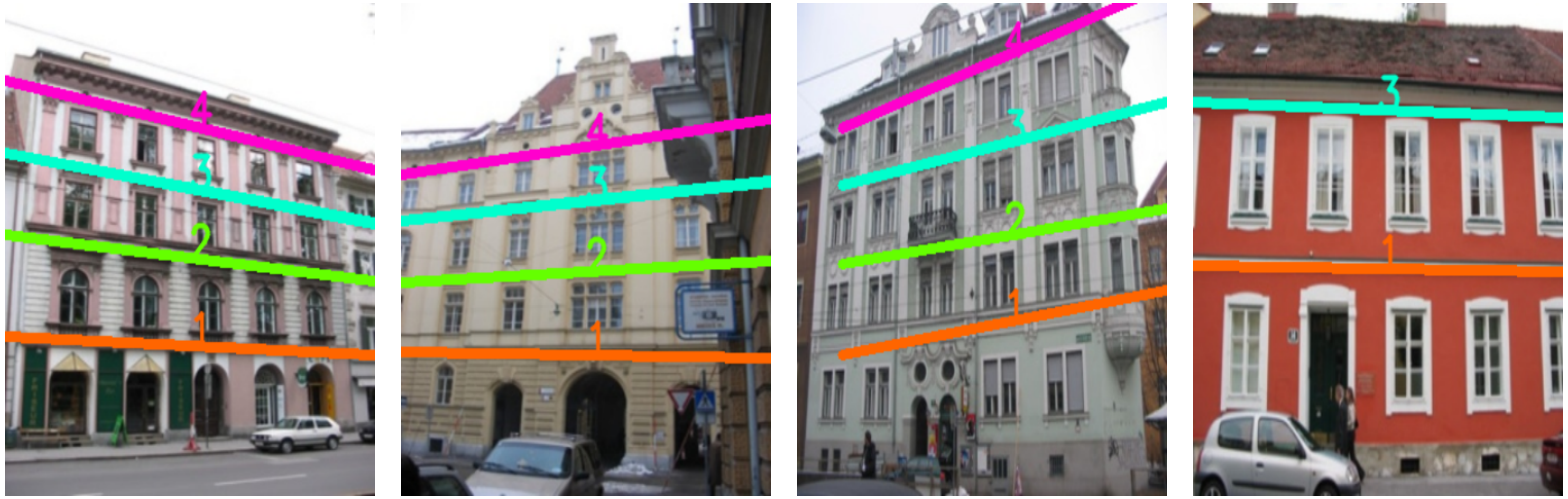
Façade parsing with Mask R-CNN



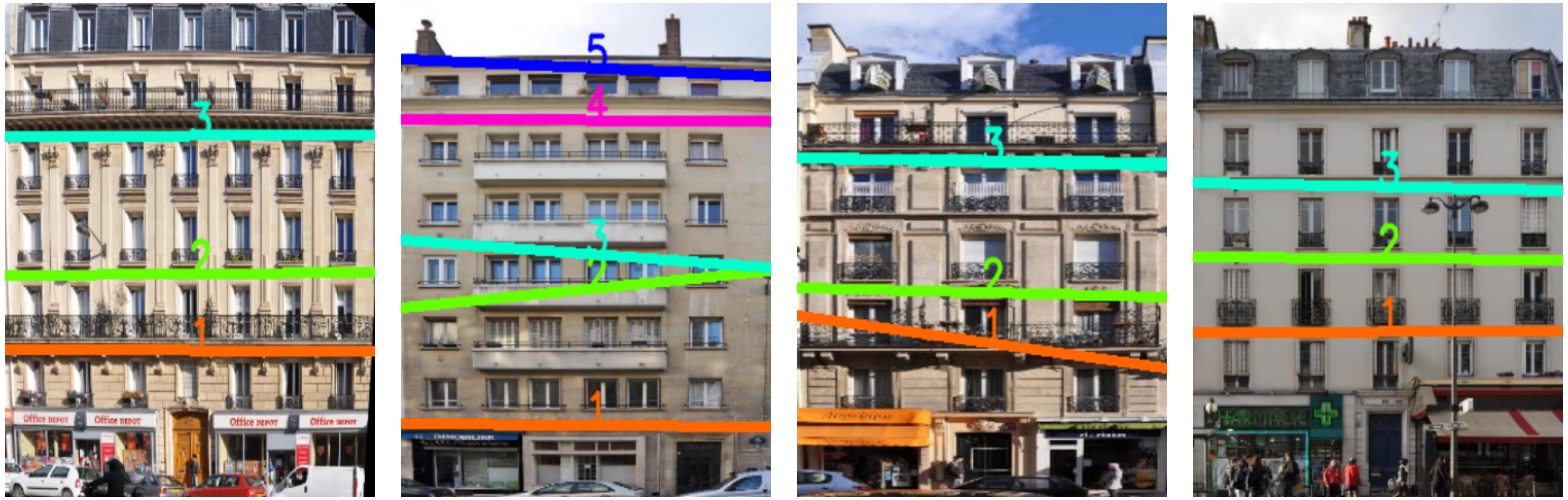
%	Detection	Segmentation
Window	71	72
Door	67	69
Sky	95	98
APs	55	57
APm	71	73
APl	95	98
AP	78	80

FloorLevel-Net

CMP:

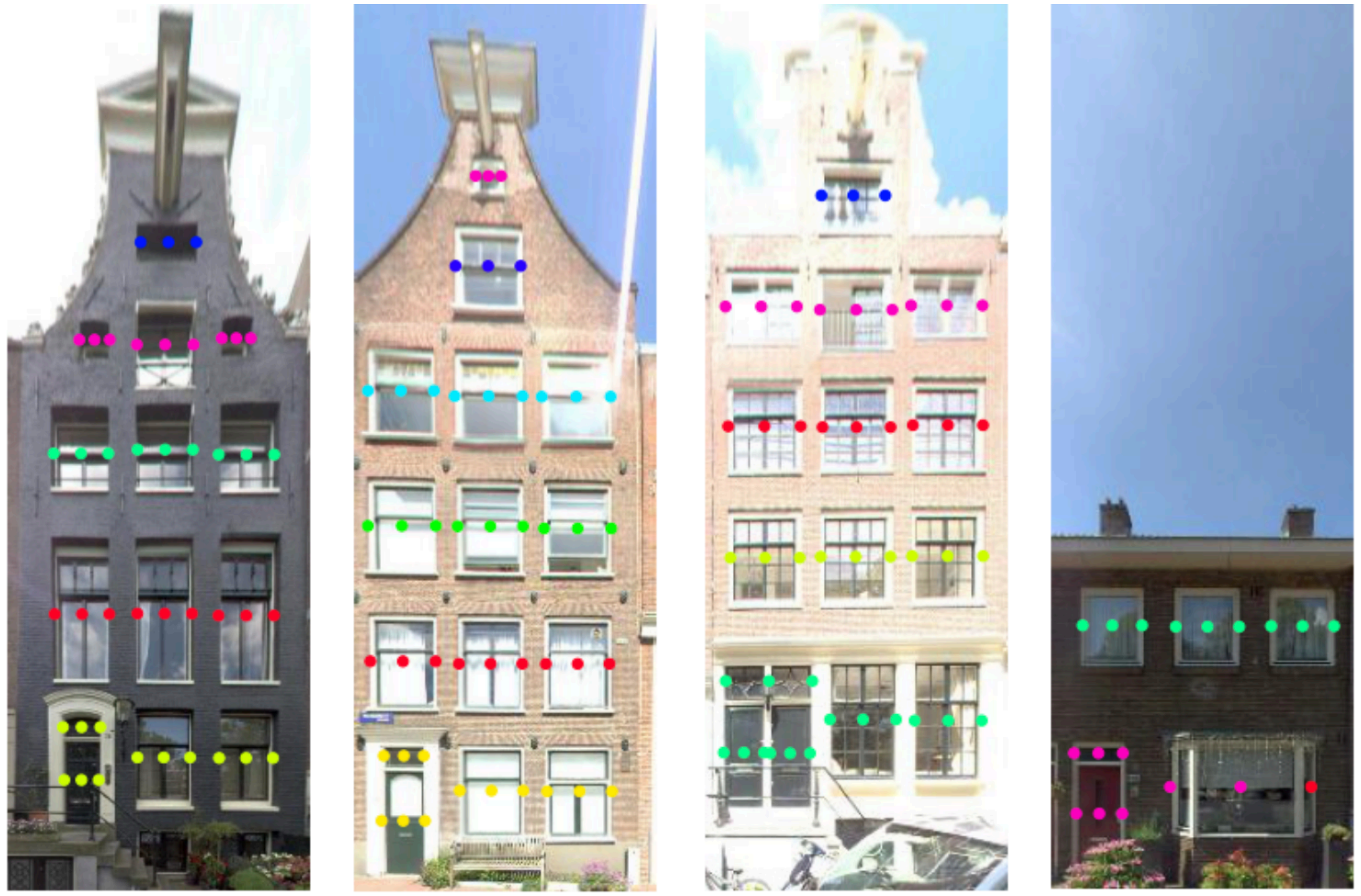


ECP:

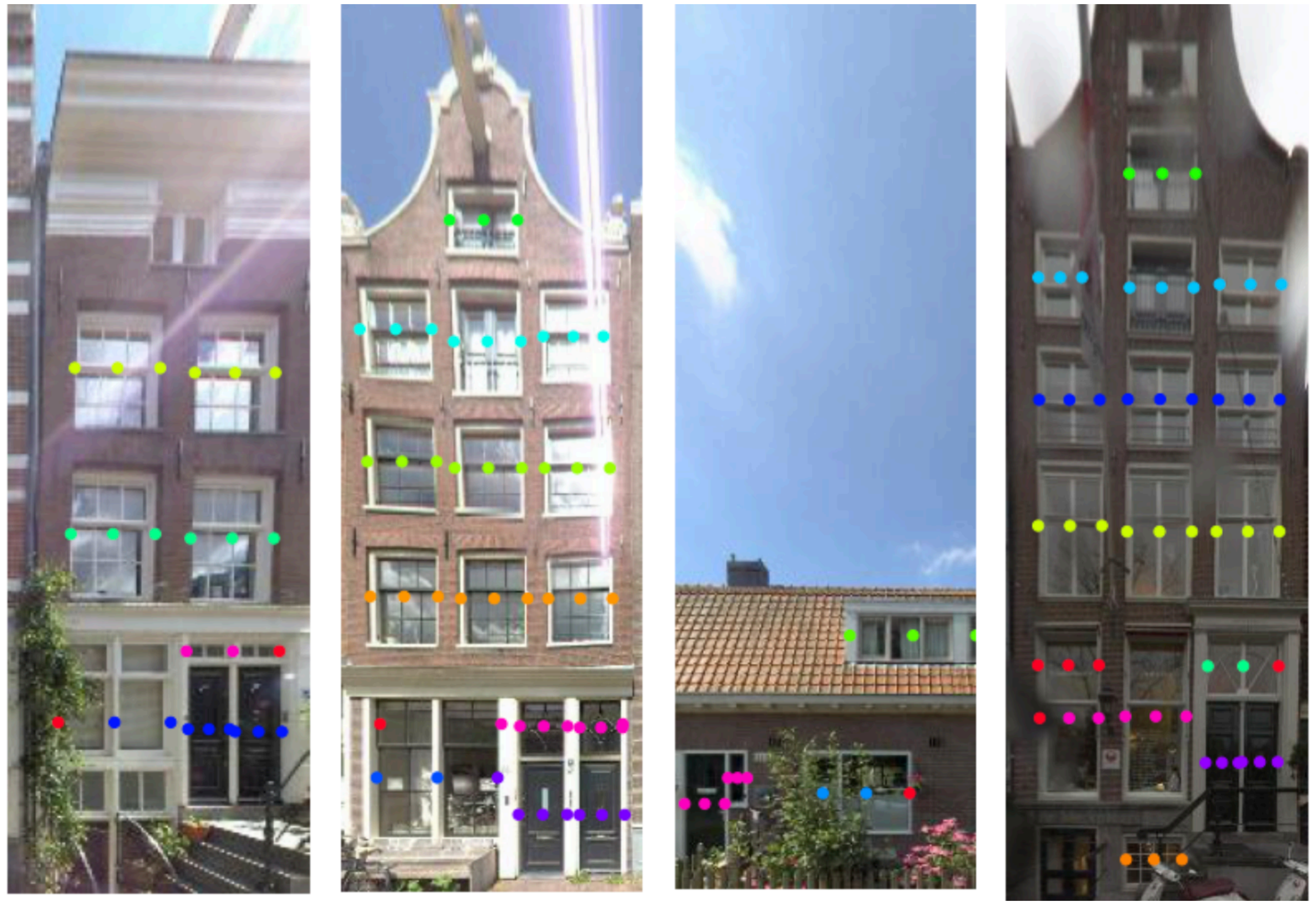


Bivariate vertical clustering

Good:



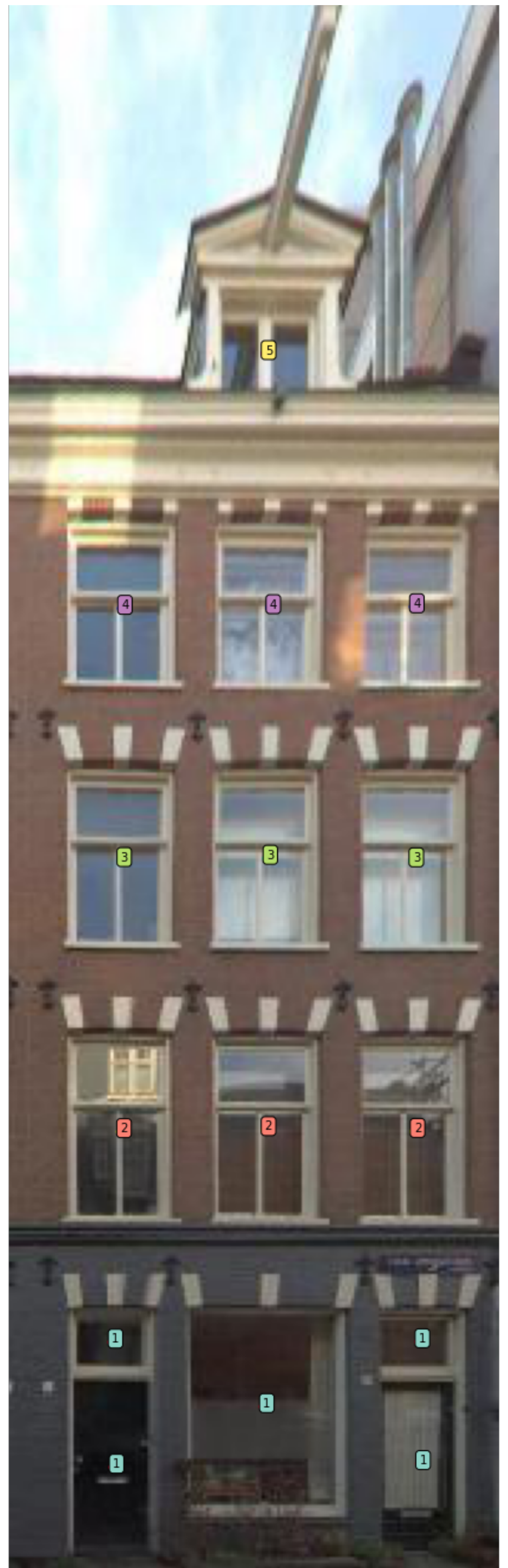
Bad:



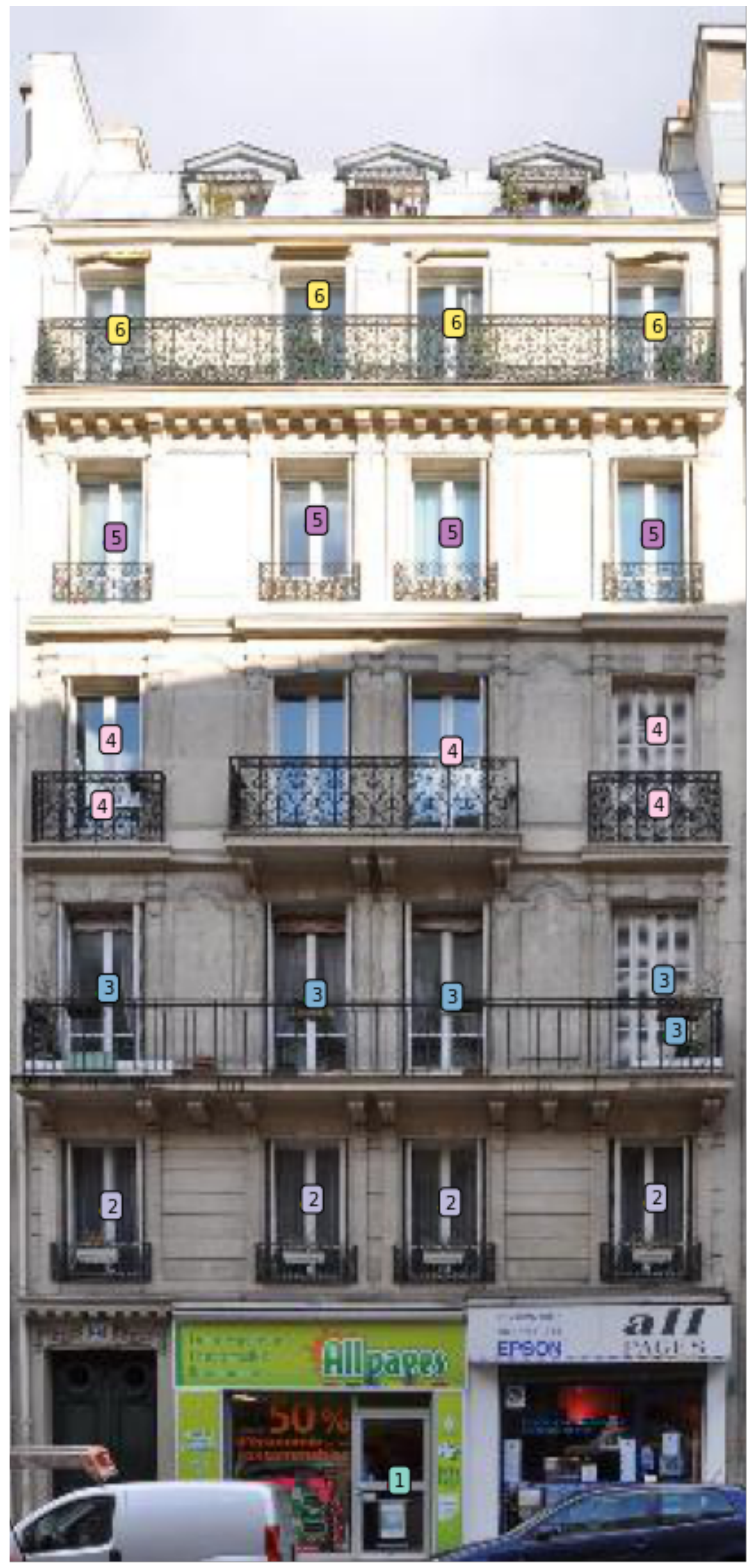
Univariate “vertical clustering”

- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

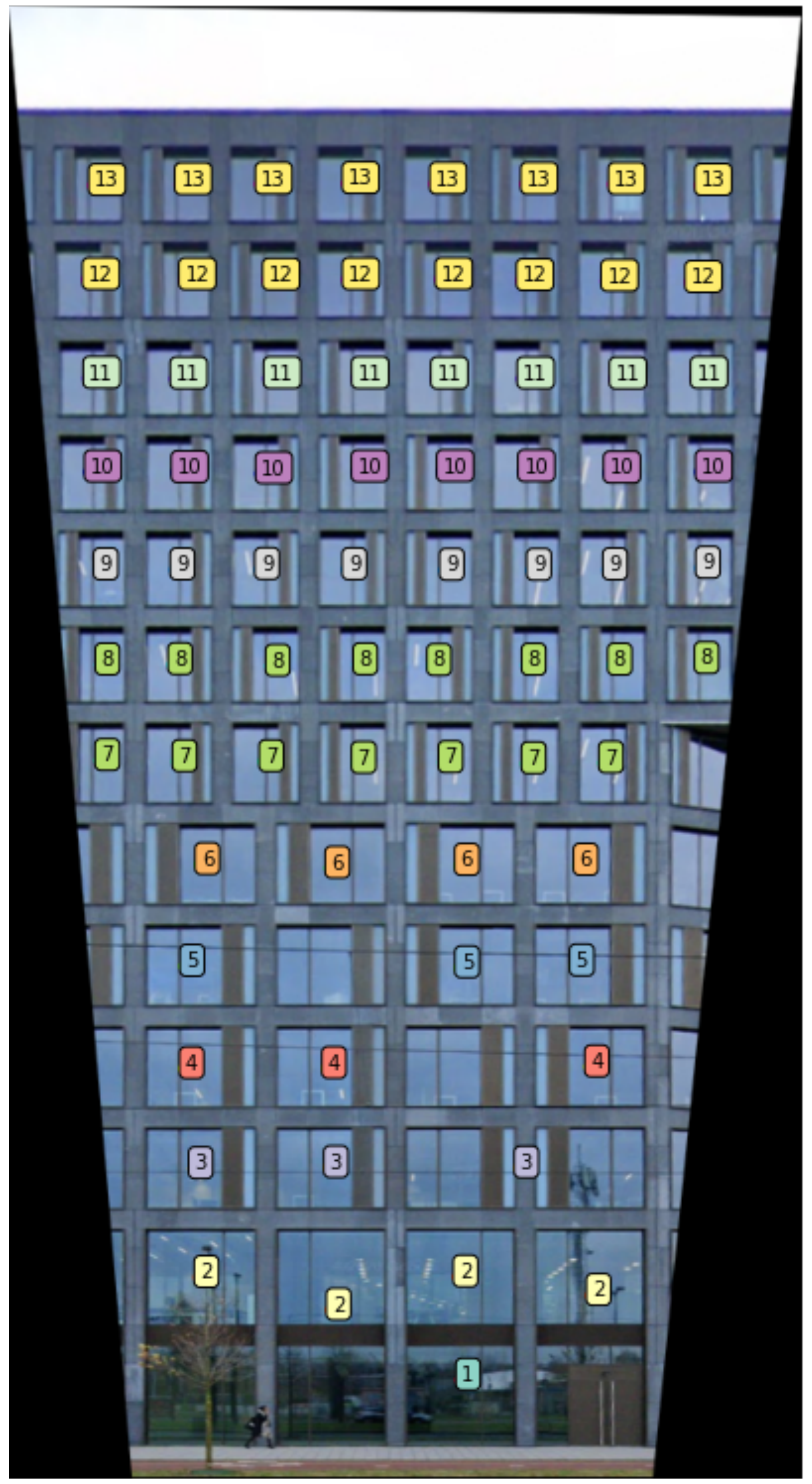
Ams. F.:



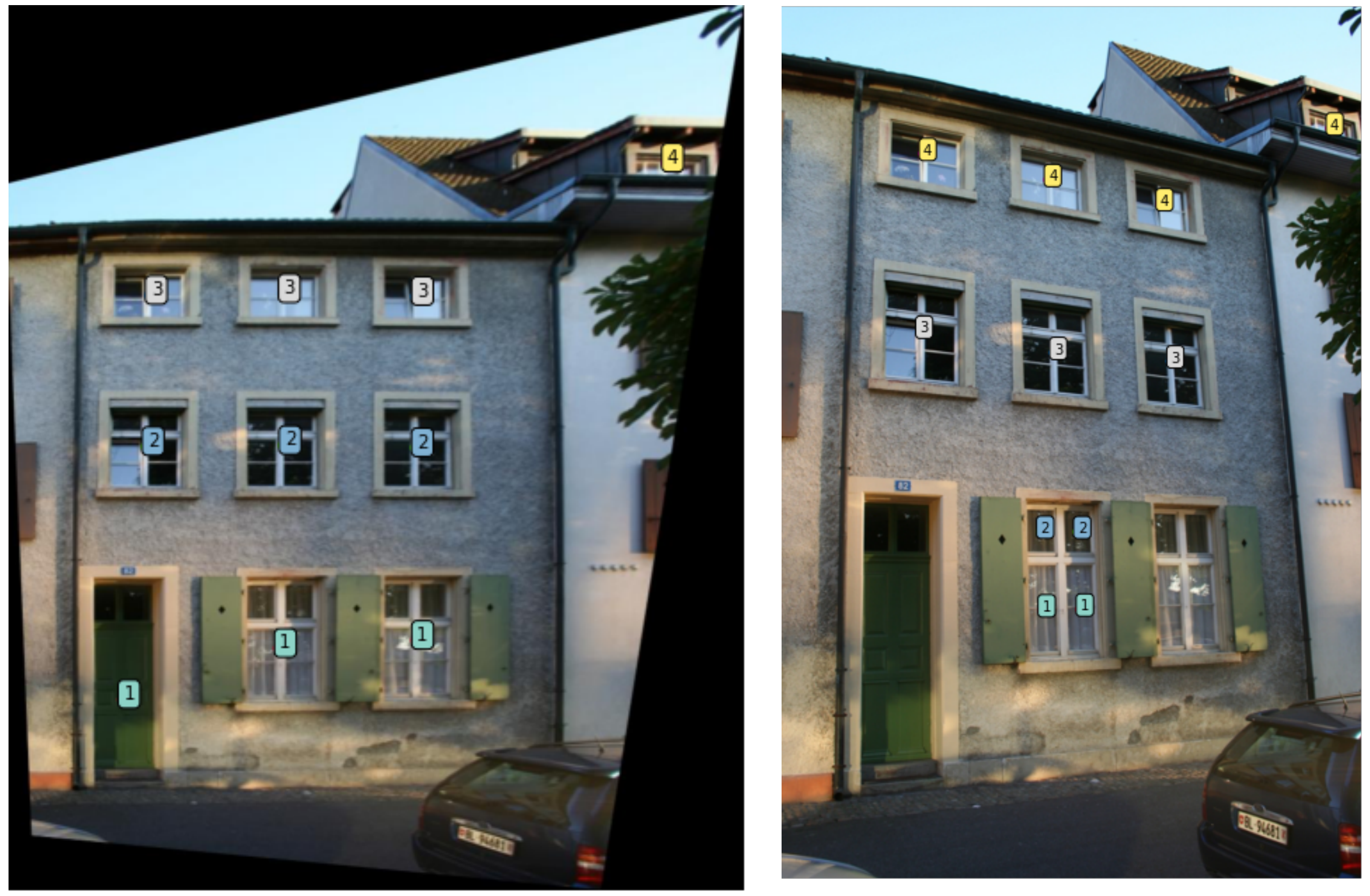
ECP:



Wild SVI:



eTRIMS:



- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

TUDelft

Univariate vertical clustering

Best results

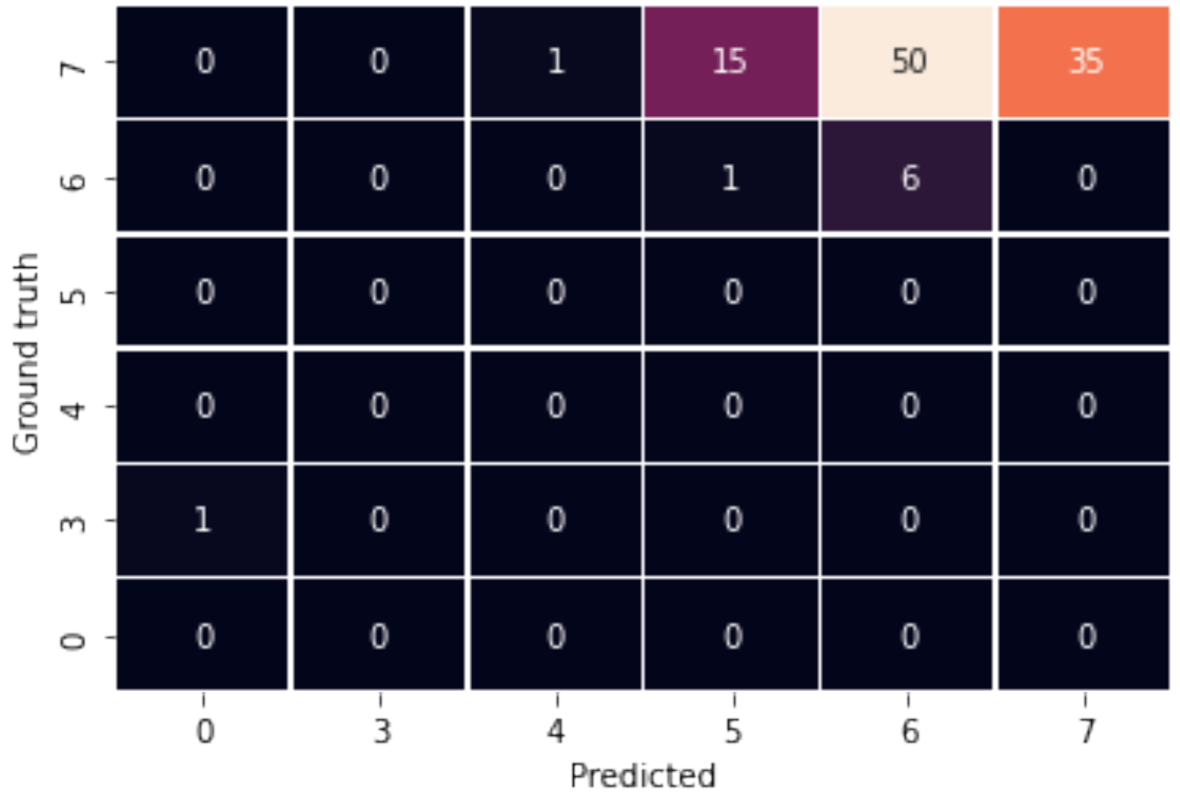
Manually tuned facade parsing model:

	Amsterdam Facade [0-7 storeys]			**Related works**		
	Detection	**Segmentation**	**Segmentation Normalised**	**Roy [13]**	**Iannelli [3]**	**Håbrekke & Nordstad [14]**
Accuracy (%) ↑	83	64	80	94.5 (<6 storeys)	85 (<5 storeys)	92
F1 (%) ↑	83	63	79			
MAE ($\mathbb{R}$) ↓	0.17	0.66	0.20			
ME ($\mathbb{R}$) ↓	-0.17	0.30	-0.20			
σ **error ($\mathbb{R}$)↓**	0.38	1.77	0.40			

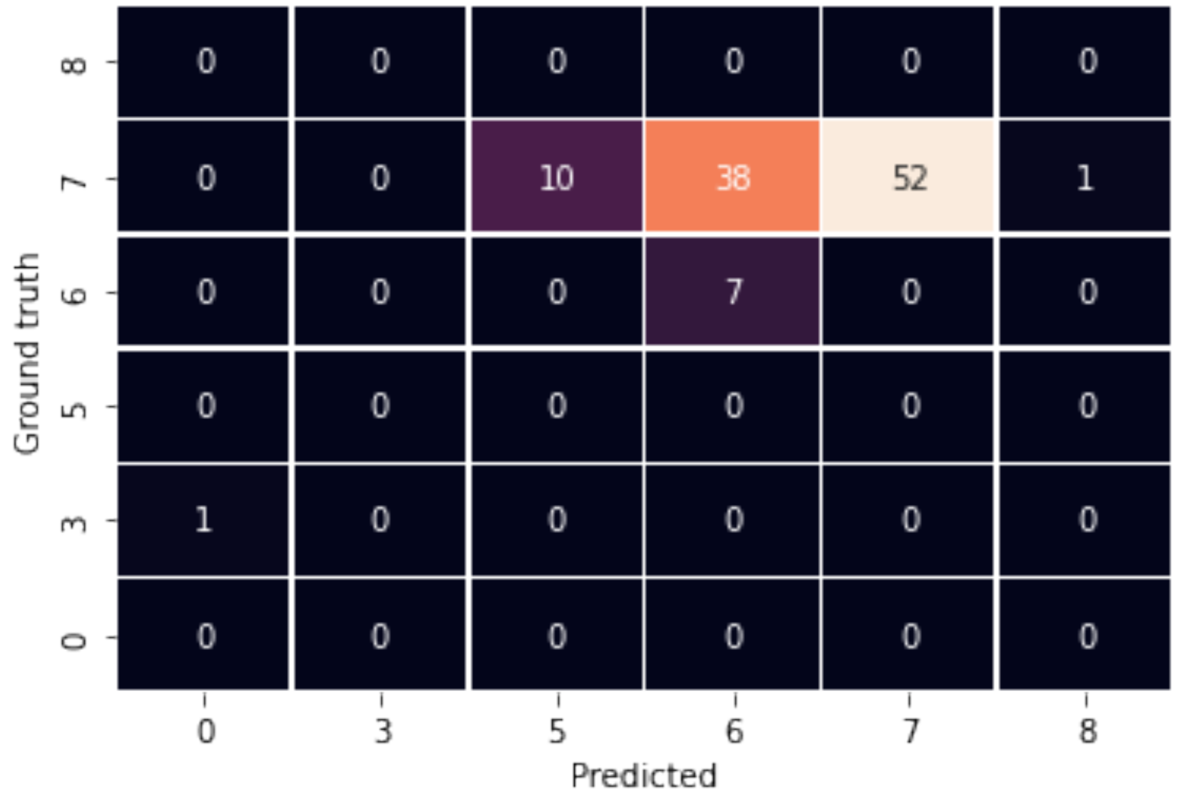
Univariate vertical clustering

Summary of evaluation on other datasets:

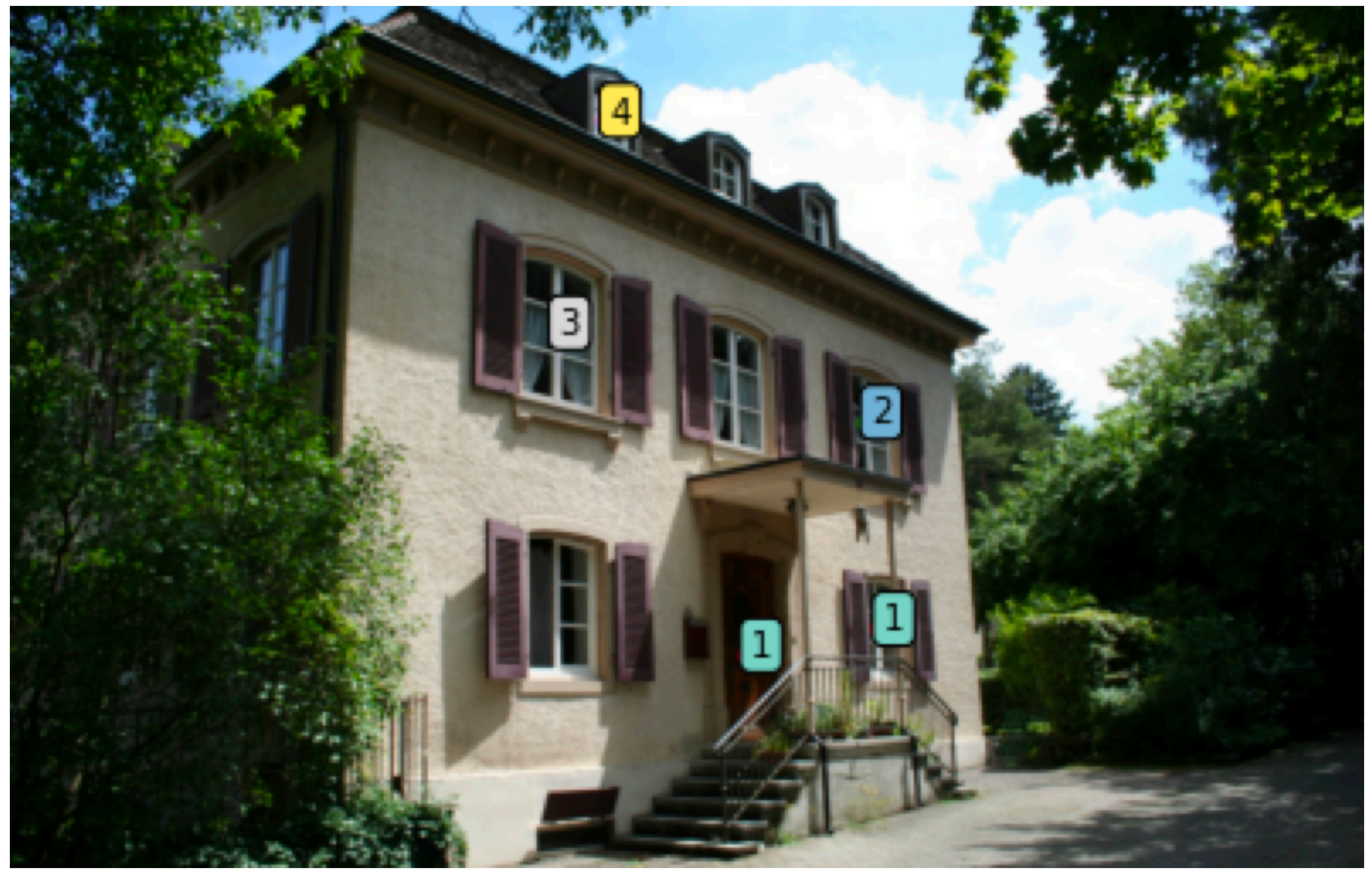
- Automatically tuned façade parsing model generalises better
- Image rectification improvement is not reflected in measurement (+/-)

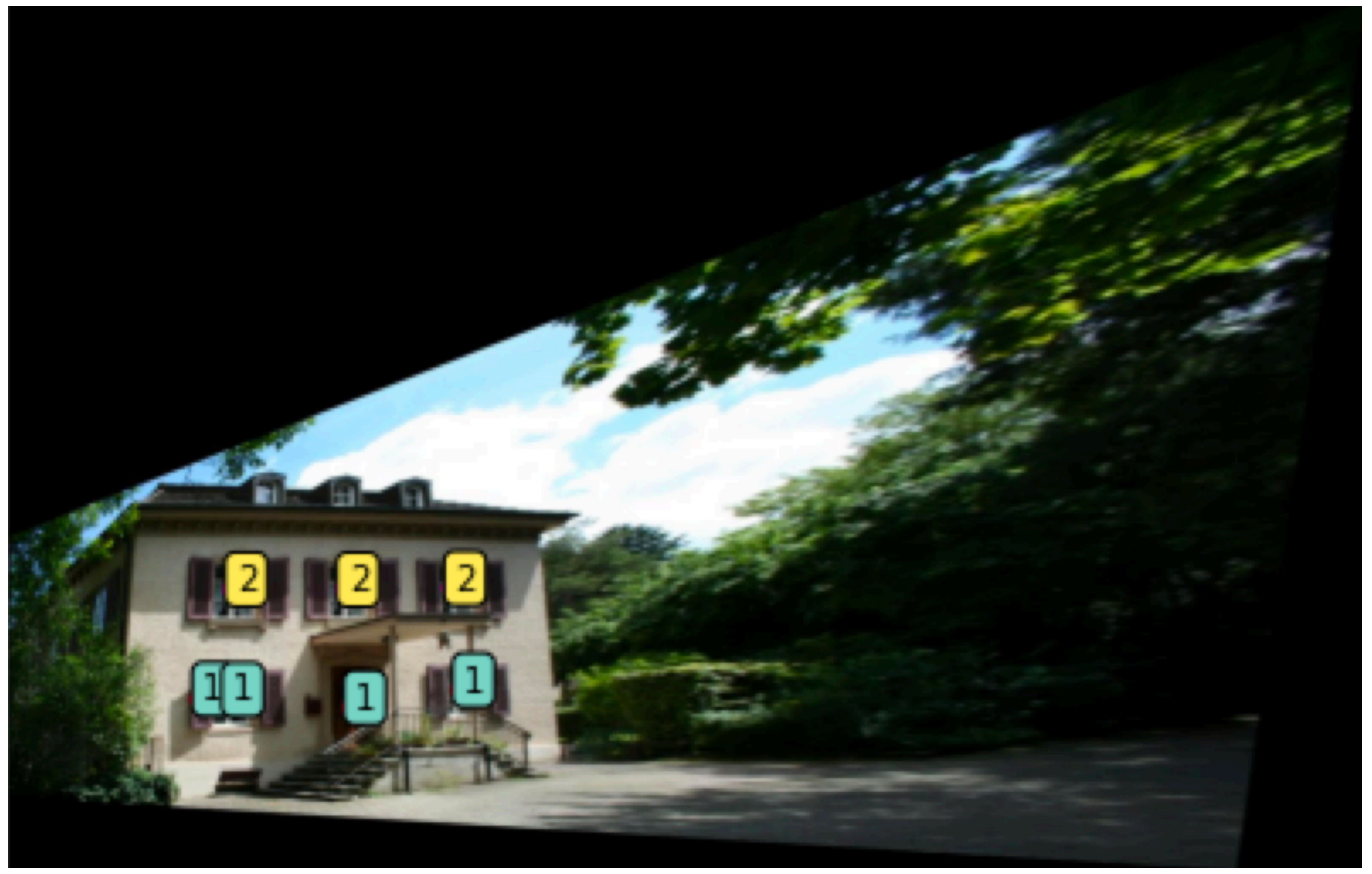


Ground truth \ Predicted	0	3	4	5	6	7
7	0	0	1	15	50	35
6	0	0	0	1	6	0
5	0	0	0	0	0	0
4	0	0	0	0	0	0
3	1	0	0	0	0	0
0	0	0	0	0	0	0



Ground truth \ Predicted	0	3	5	6	7	8
8	0	0	0	0	0	0
7	0	0	10	38	52	1
6	0	0	0	7	0	0
5	0	0	0	0	0	0
3	1	0	0	0	0	0
0	0	0	0	0	0	0



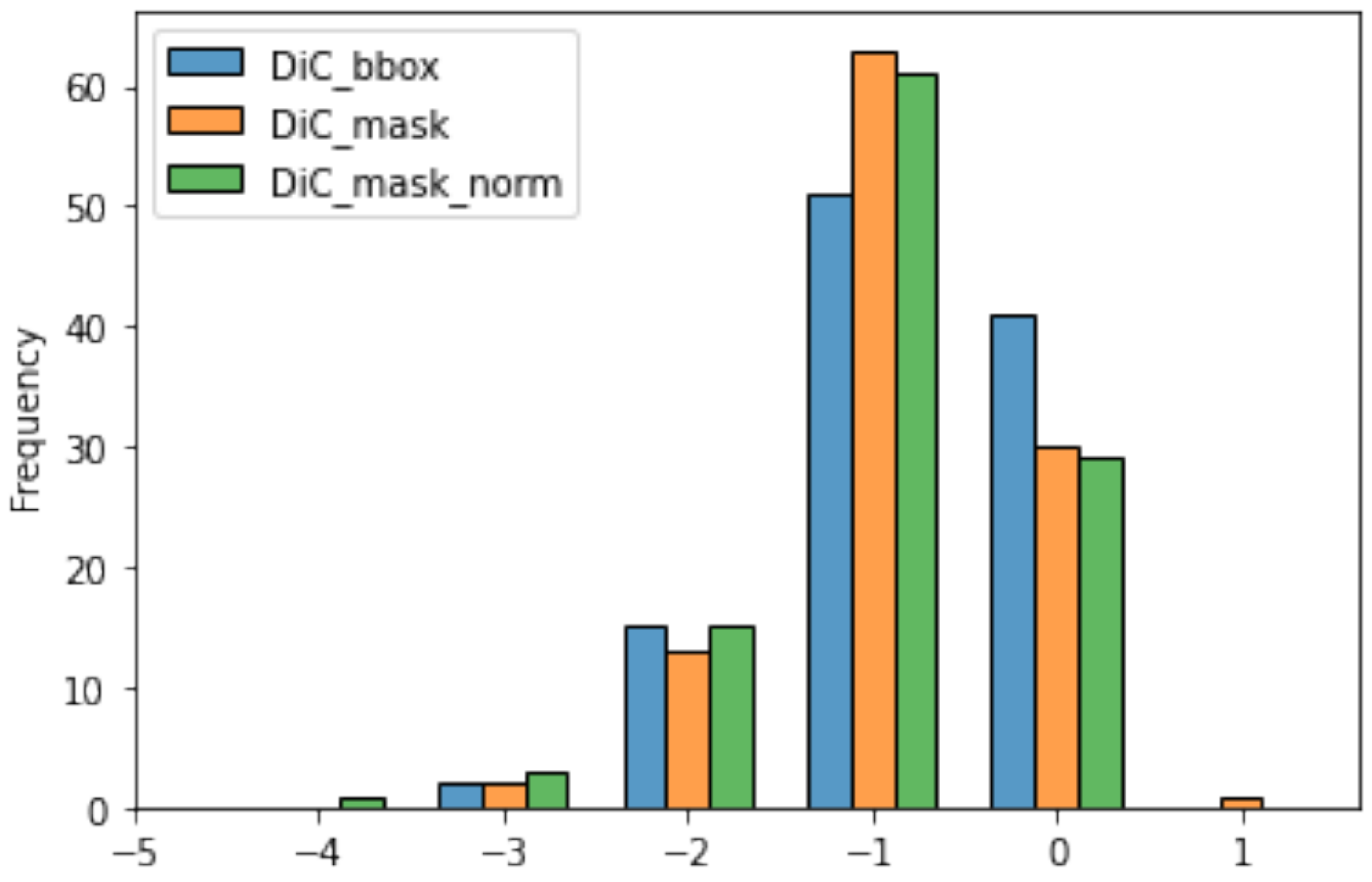


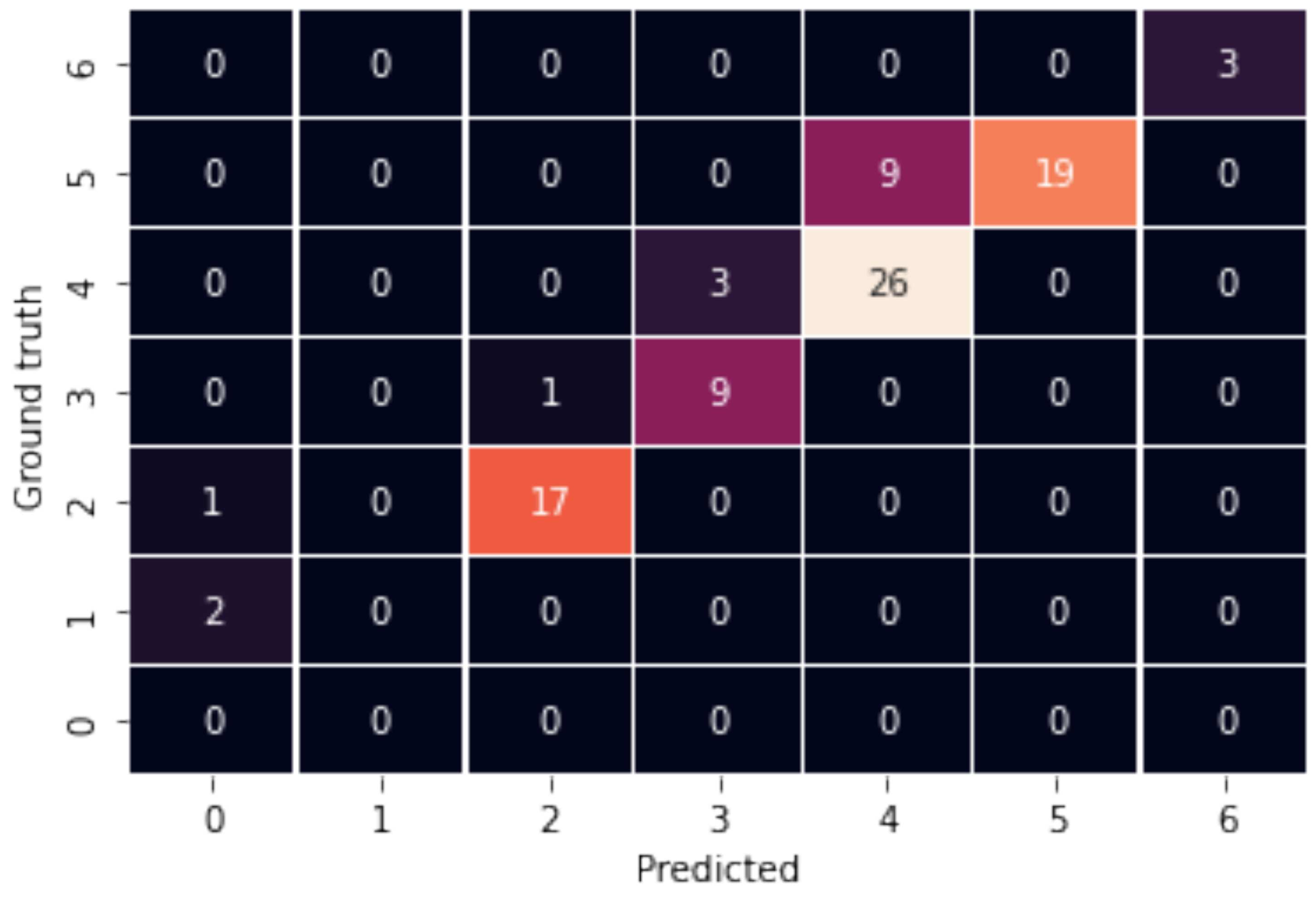
TU Delft

- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

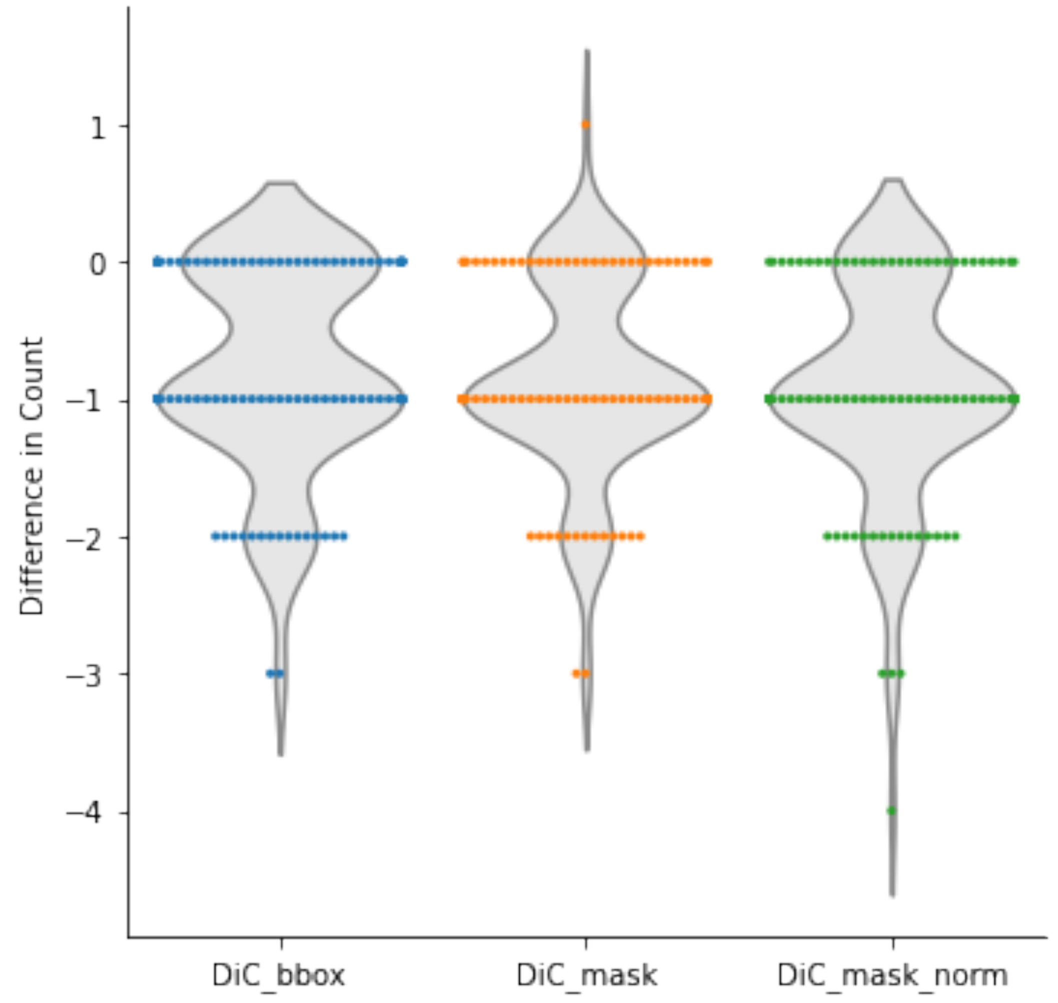
TUDelft

Undershooting





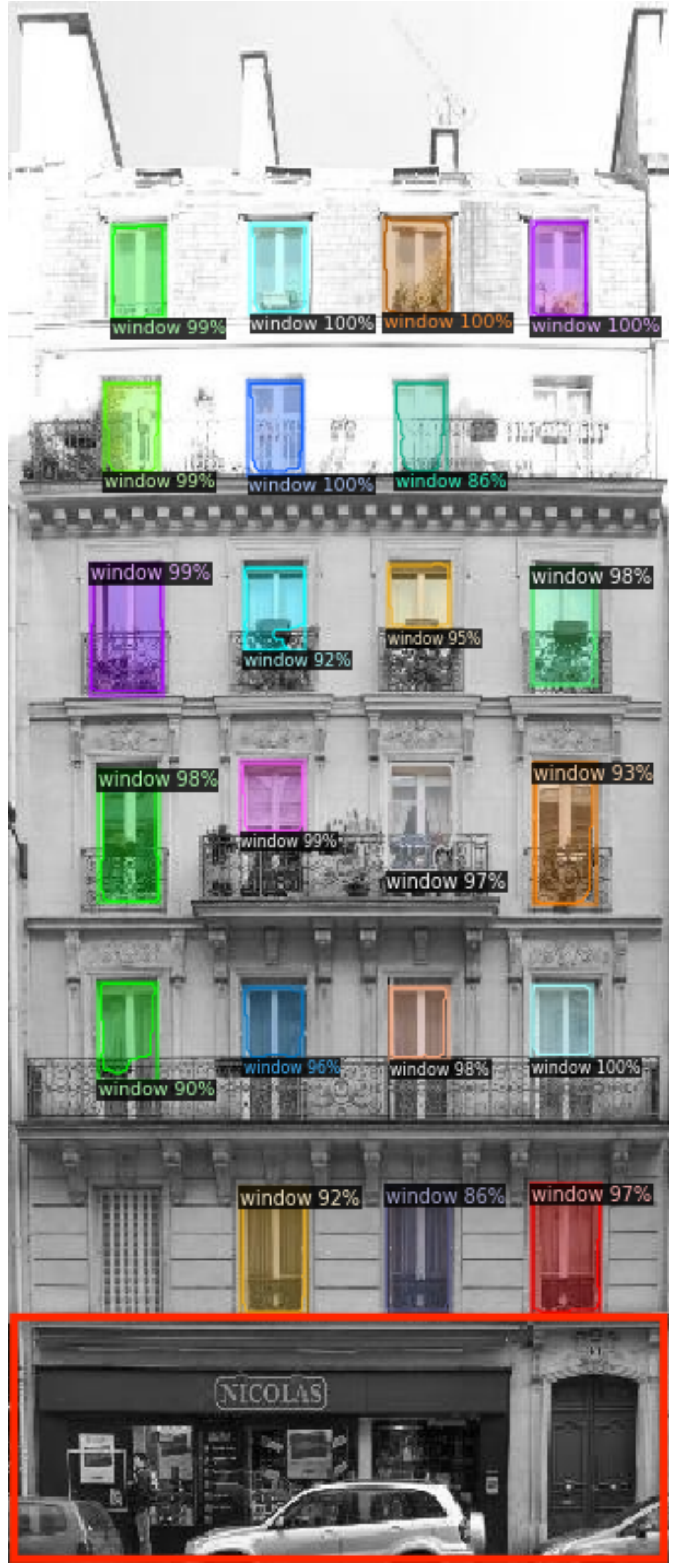
Ground truth \ Predicted	0	1	2	3	4	5	6
6	0	0	0	0	0	0	3
5	0	0	0	0	9	19	0
4	0	0	0	3	26	0	0
3	0	0	1	9	0	0	0
2	1	0	17	0	0	0	0
1	2	0	0	0	0	0	0
0	0	0	0	0	0	0	0

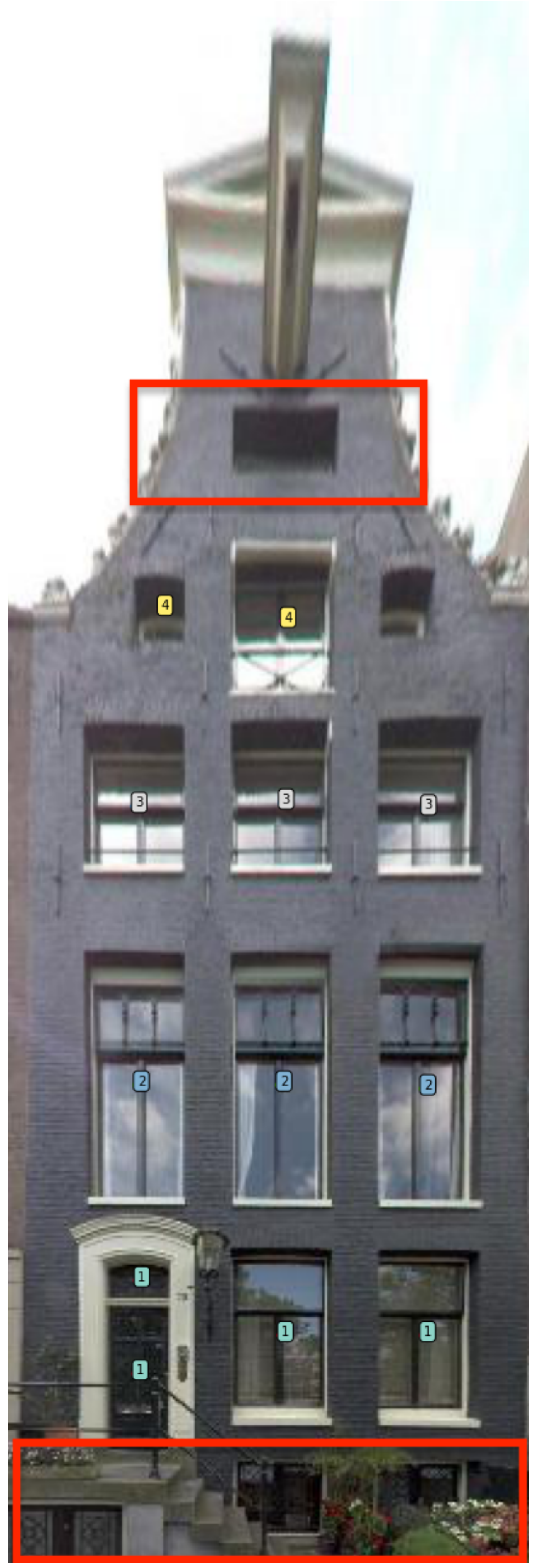


- Motivation
- Related work
- Background
- Objectives
- Methodology
- Experiments
- Results
- Conclusions

TU Delft

Undershooting

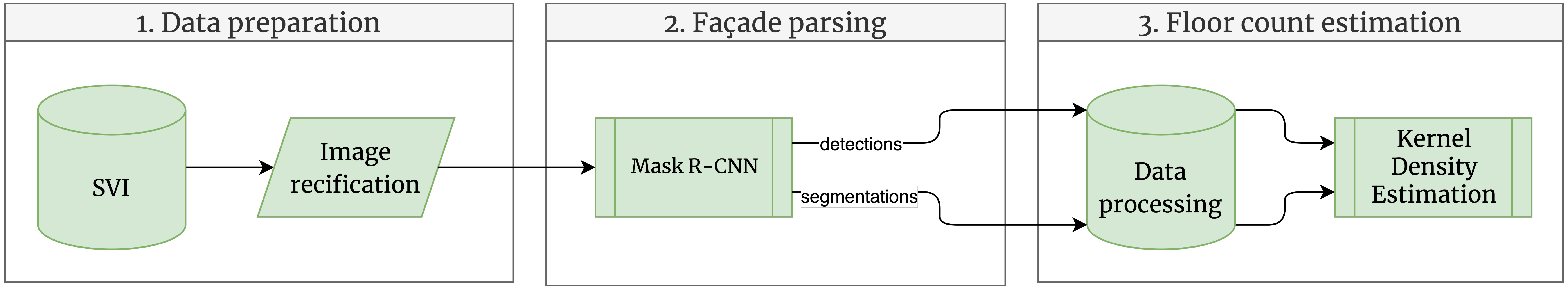




Conclusions

Conclusions

How to determine floor count in an image with the use of learning-based façade parsing?



1. Data preparation

SVI

Image recification

2. Façade parsing

Mask R-CNN

detections

segmentations

3. Floor count estimation

Data processing

Kernel Density Estimation

- **Promising results** for small scale, considering no discrimination in storey-numbers
- **Mask R-CNN** for façade parsing works well, also gives opportunity to have both detections and segmentations
  - Improvement in façade parsing performance can:
    - Overcome undershooting
    - More robust in rectified SVI
  - Automatically tuned façade parsing model most versatile
- **Data processing**: Detections -> point selection. Segmentations -> bitmap to pixel-coordinates
- **KDE**, with maxima finding works well. Combine manual + automatic tuning

Limitations

- **Dataset**: lack in variability, ground-truth availability, annotation quality
- **Breadth of research**: jack of all trades, master of none
- **SVI coverage and practicality**: simplification of problem, no use of API
- **Computation limitations**: Conservative training routines employed

Future work

- **Dataset creation**: use of API, open-source, variability, ground-truth availability, annotation quality, automatic façade retrieval
- **Model sophistication**: FLN —> training for higher level semantics, use of attention modules. Also, increase speed.
- **Literature review**: floor count standards, regulations, exception cases
- **Improve vertical clustering**: KDE optimisation, eg parameter search, manual + automatic harmonisation

Thank you for listening!

Any questions?

References

[1] By Eugen Simion 14 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=45823854

[2] Biljecki, F. (2017). Level of detail in 3D city models.

[3] Iannelli, G. C., & Dell'Acqua, F. (2017). Extensive exposure mapping in urban areas through deep analysis of street-level pictures for floor count determination. Urban Science, 1(2), 16.

[4] Rosenfelder, M., Wussow, M., Gust, G., Cremades, R., and Neumann, D. (2021). Predicting residential electricity consumption using aerial and street view images. Applied Energy, 301:117407.

[5] Chen, F.-C., Subedi, A., Jahanshahi, M. R., Johnson, D. R., and Delp, E. J. (2022). Deep learning--based building attribute estimation from google street view images for flood risk assessment using feature fusion and task relation encoding. Journal of Computing in Civil Engineering, 36(6):04022031.

[6] Sun, Y., Malihi, S., Li, H., and Maboudi, M. (2022). Deepwindows: Windows instance segmentation through an improved mask r-cnn using spatial attention and relation modules. ISPRS International Journal of Geo-Information, 11(3):162.

[7] Casado-García, Á., Domínguez, C., García-Domínguez, M. et al. CLoDSA: a tool for augmentation in classification, localization, detection, semantic segmentation and instance segmentation tasks. BMC Bioinformatics 20, 323 (2019). https://doi.org/10.1186/12859-019-2931-1

[8] Elgendy, M. (2020). Deep learning for vision systems. Simon and Schuster.

[9] Liu, Hantang, et al. "Deepfacade: A deep learning approach to facade parsing." IJCAI, 2017.

[10] Tylecek, R. and Sára, R. (2010). A weak structure model for regular pattern recognition applied to facade images. In Asian Conference on Computer Vision, pages 450-463. Springer.

[11] Ayenew, M. (2021). Towards large scale façade parsing: A deep learning pipeline using mask r-cnn.

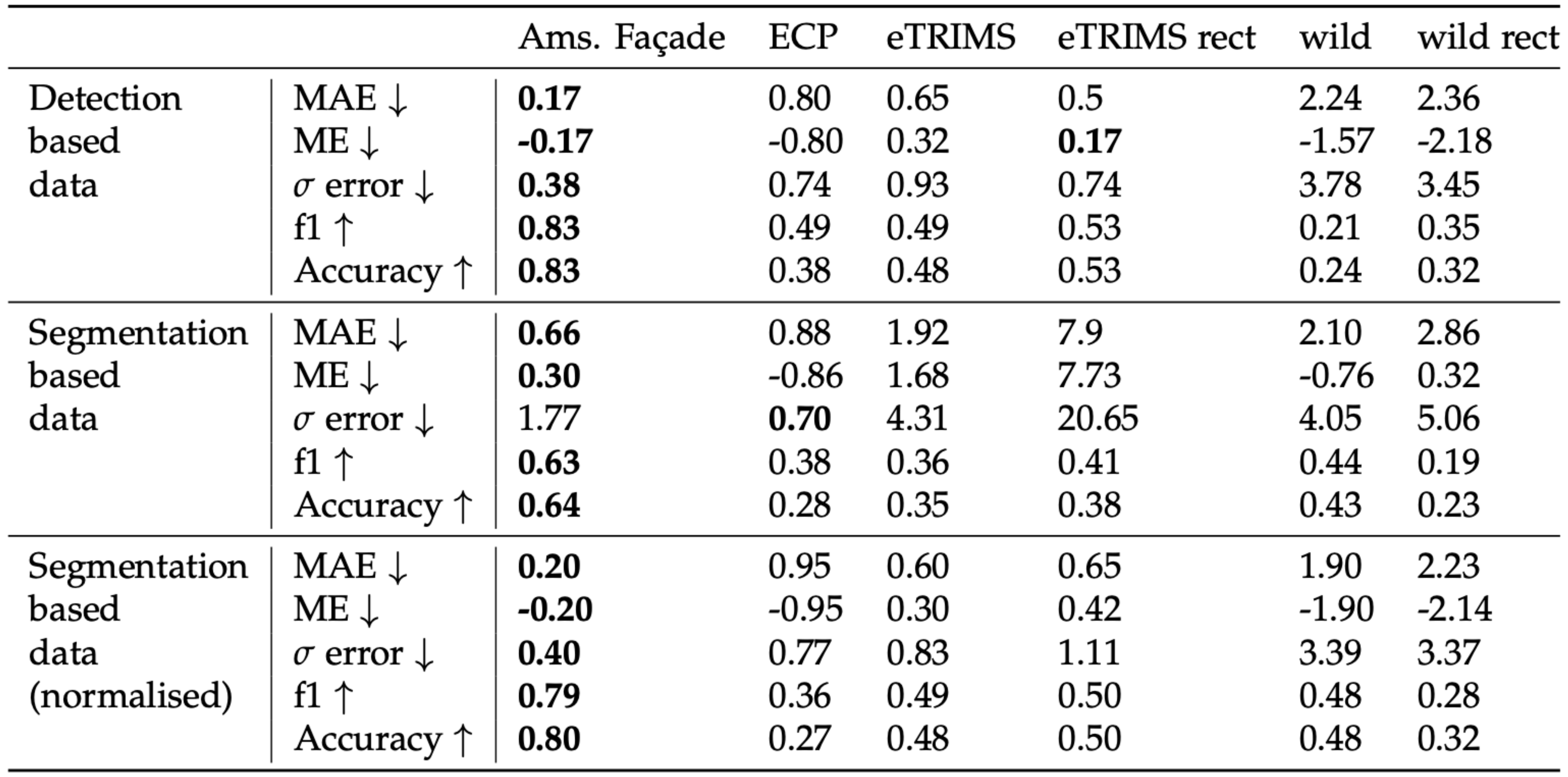
[12] Wu, M., Zeng, W., and Fu, C.-W. (2021). Floorlevel-net: Recognizing floor-level lines with height-attention-guided multi-task learning. IEEE Transactions on Image Processing, 30:6686-6699.

[13] Roy, E. (2022). Inferring the number of floors of building footprints in the netherlands. Master's thesis, Delft University of Technology.

[14] Håbrekke, and Nordstad, F. D. (2022). Estimating the height of facades with street-level imagery using facade parsing, floor segmentation, and urban rules. Master's thesis, Nor- wegian University of Science and Technology.

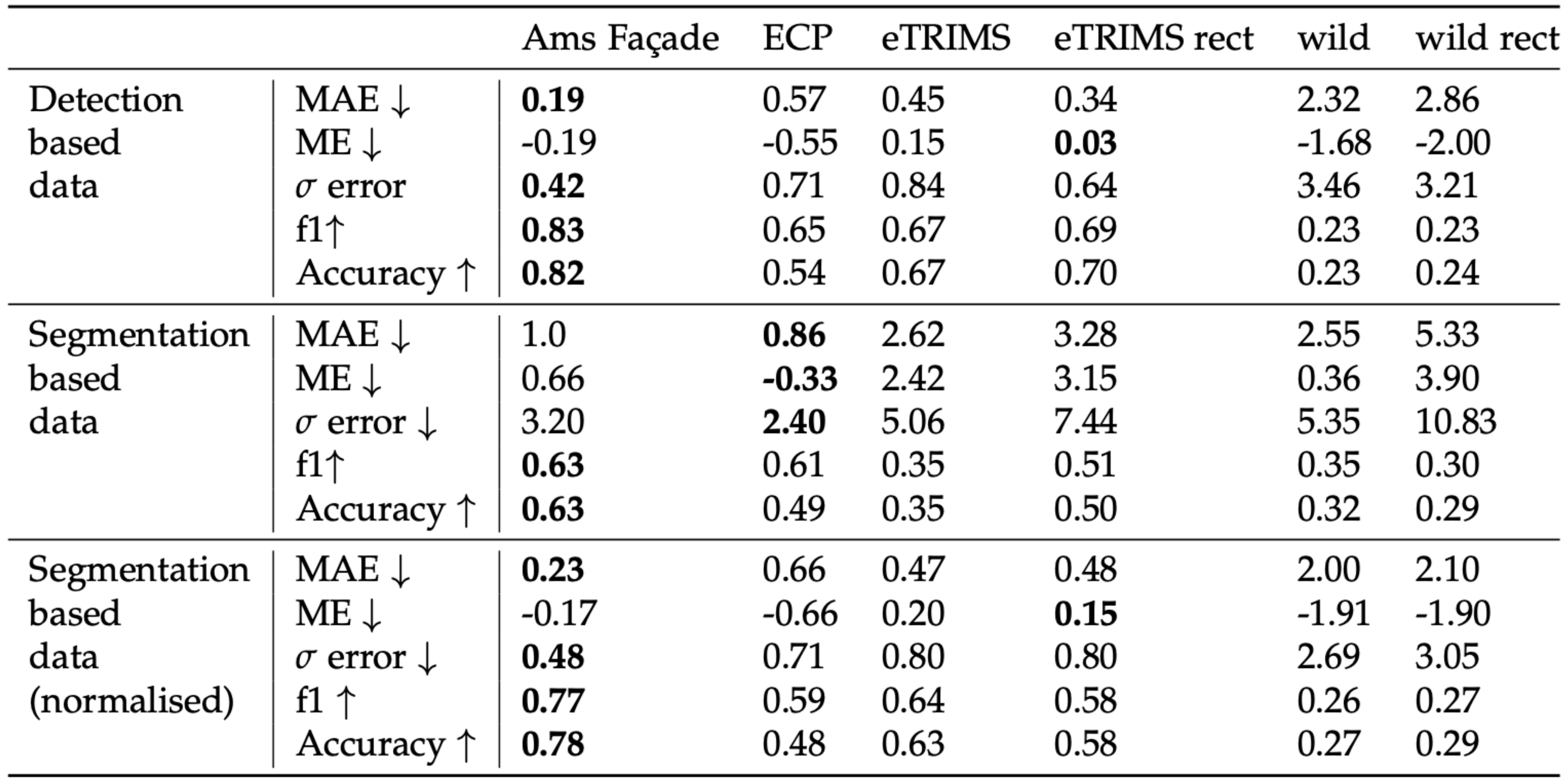
Additional results

Manually tuned facade parsing model:



		Ams. Façade	ECP	eTRIMS	eTRIMS rect	wild	wild rect
Detection based data	MAE ↓	**0.17**	0.80	0.65	0.5	2.24	2.36
	ME ↓	**-0.17**	-0.80	0.32	**0.17**	-1.57	-2.18
	σ error ↓	**0.38**	0.74	0.93	0.74	3.78	3.45
	f1 ↑	**0.83**	0.49	0.49	0.53	0.21	0.35
	Accuracy ↑	**0.83**	0.38	0.48	0.53	0.24	0.32
Segmentation based data	MAE ↓	**0.66**	0.88	1.92	7.9	2.10	2.86
	ME ↓	**0.30**	-0.86	1.68	7.73	-0.76	0.32
	σ error ↓	1.77	**0.70**	4.31	20.65	4.05	5.06
	f1 ↑	**0.63**	0.38	0.36	0.41	0.44	0.19
	Accuracy ↑	**0.64**	0.28	0.35	0.38	0.43	0.23
Segmentation based data (normalised)	MAE ↓	**0.20**	0.95	0.60	0.65	1.90	2.23
	ME ↓	**-0.20**	-0.95	0.30	0.42	-1.90	-2.14
	σ error ↓	**0.40**	0.77	0.83	1.11	3.39	3.37
	f1 ↑	**0.79**	0.36	0.49	0.50	0.48	0.28
	Accuracy ↑	**0.80**	0.27	0.48	0.50	0.48	0.32

Automatically tuned facade parsing model:



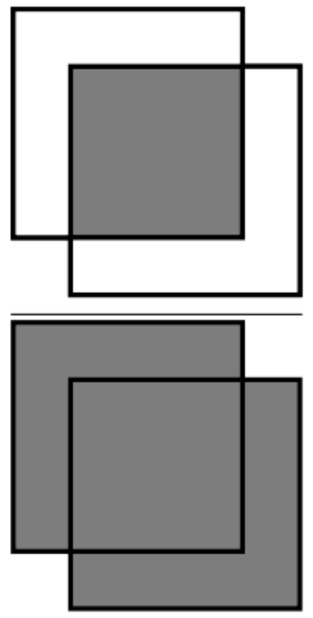
		Ams Façade	ECP	eTRIMS	eTRIMS rect	wild	wild rect
Detection based data	MAE ↓	**0.19**	0.57	0.45	0.34	2.32	2.86
	ME ↓	-0.19	-0.55	0.15	**0.03**	-1.68	-2.00
	σ error	**0.42**	0.71	0.84	0.64	3.46	3.21
	f1↑	**0.83**	0.65	0.67	0.69	0.23	0.23
	Accuracy ↑	**0.82**	0.54	0.67	0.70	0.23	0.24
Segmentation based data	MAE ↓	1.0	**0.86**	2.62	3.28	2.55	5.33
	ME ↓	0.66	**-0.33**	2.42	3.15	0.36	3.90
	σ error ↓	3.20	**2.40**	5.06	7.44	5.35	10.83
	f1↑	**0.63**	0.61	0.35	0.51	0.35	0.30
	Accuracy ↑	**0.63**	0.49	0.35	0.50	0.32	0.29
Segmentation based data (normalised)	MAE ↓	**0.23**	0.66	0.47	0.48	2.00	2.10
	ME ↓	-0.17	-0.66	0.20	**0.15**	-1.91	-1.90
	σ error ↓	**0.48**	0.71	0.80	0.80	2.69	3.05
	f1 ↑	**0.77**	0.59	0.64	0.58	0.26	0.27
	Accuracy ↑	**0.78**	0.48	0.63	0.58	0.27	0.29

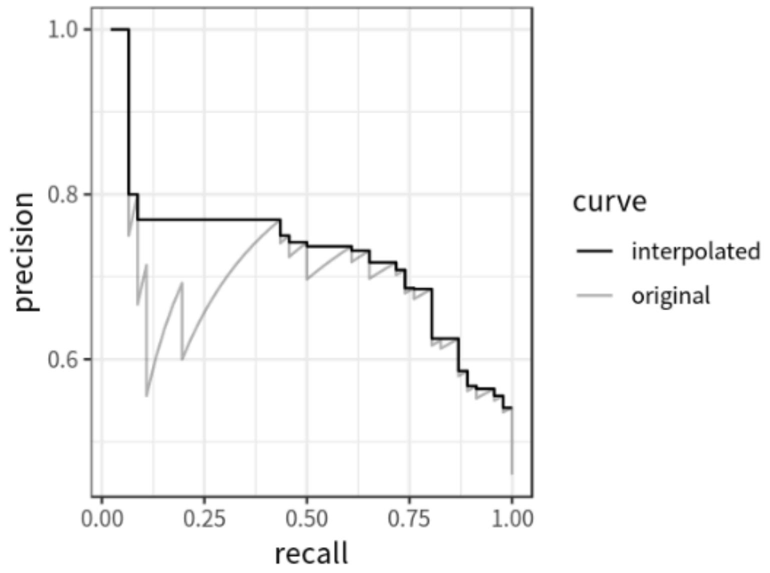
Metrics

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$IoU = \frac{\text{area of overlap}}{\text{area of union}} =$$





$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n} 1(y_i = \hat{y}_i)$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

$$\text{ME} = \frac{\sum_{i=1}^{n} y_i - \hat{y}_i}{n}$$

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

$$DiC = \#L^{pred} - \#L^{gt}$$

Motivation

Related work

Background

Objectives

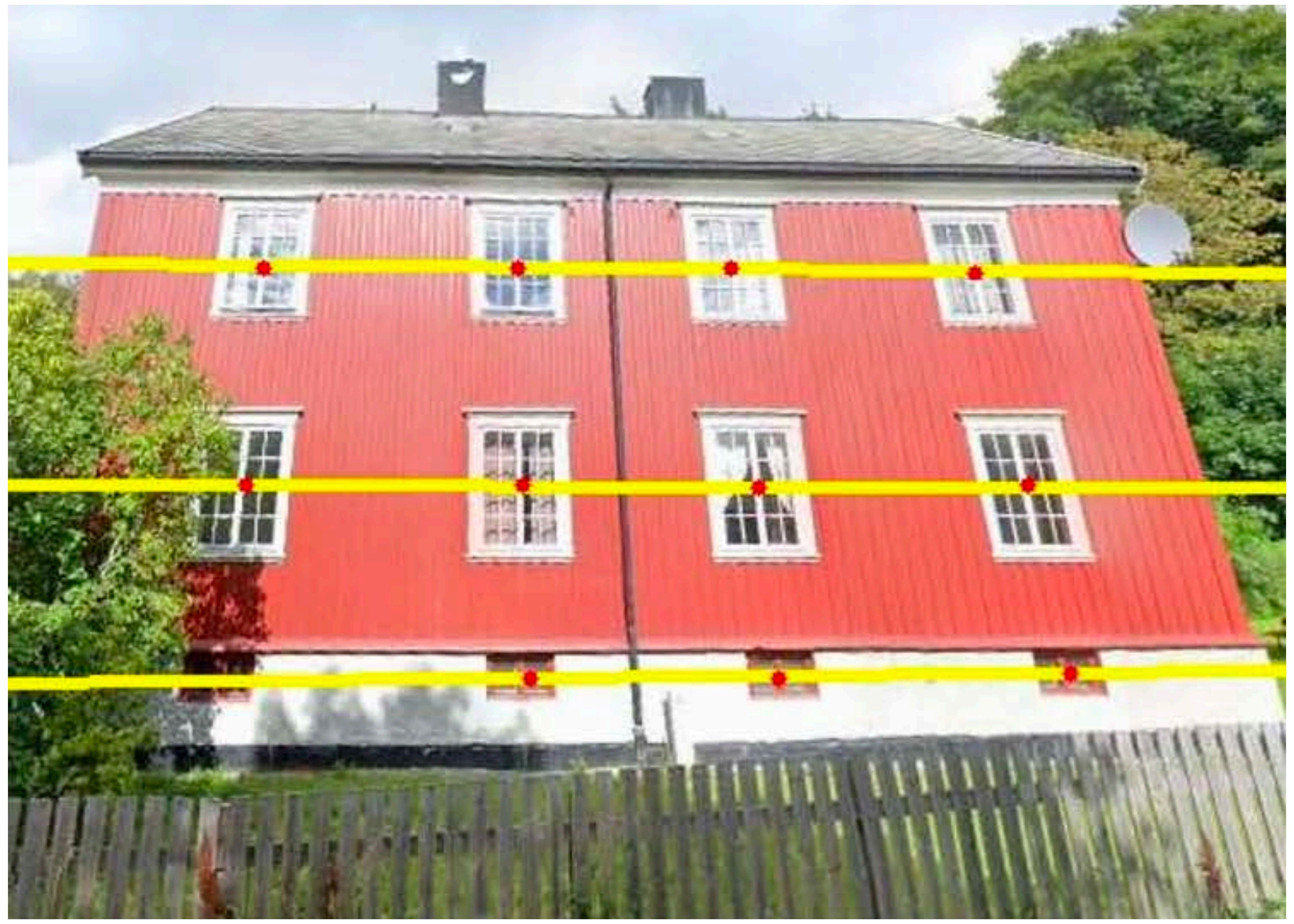
Methodology

Experiments

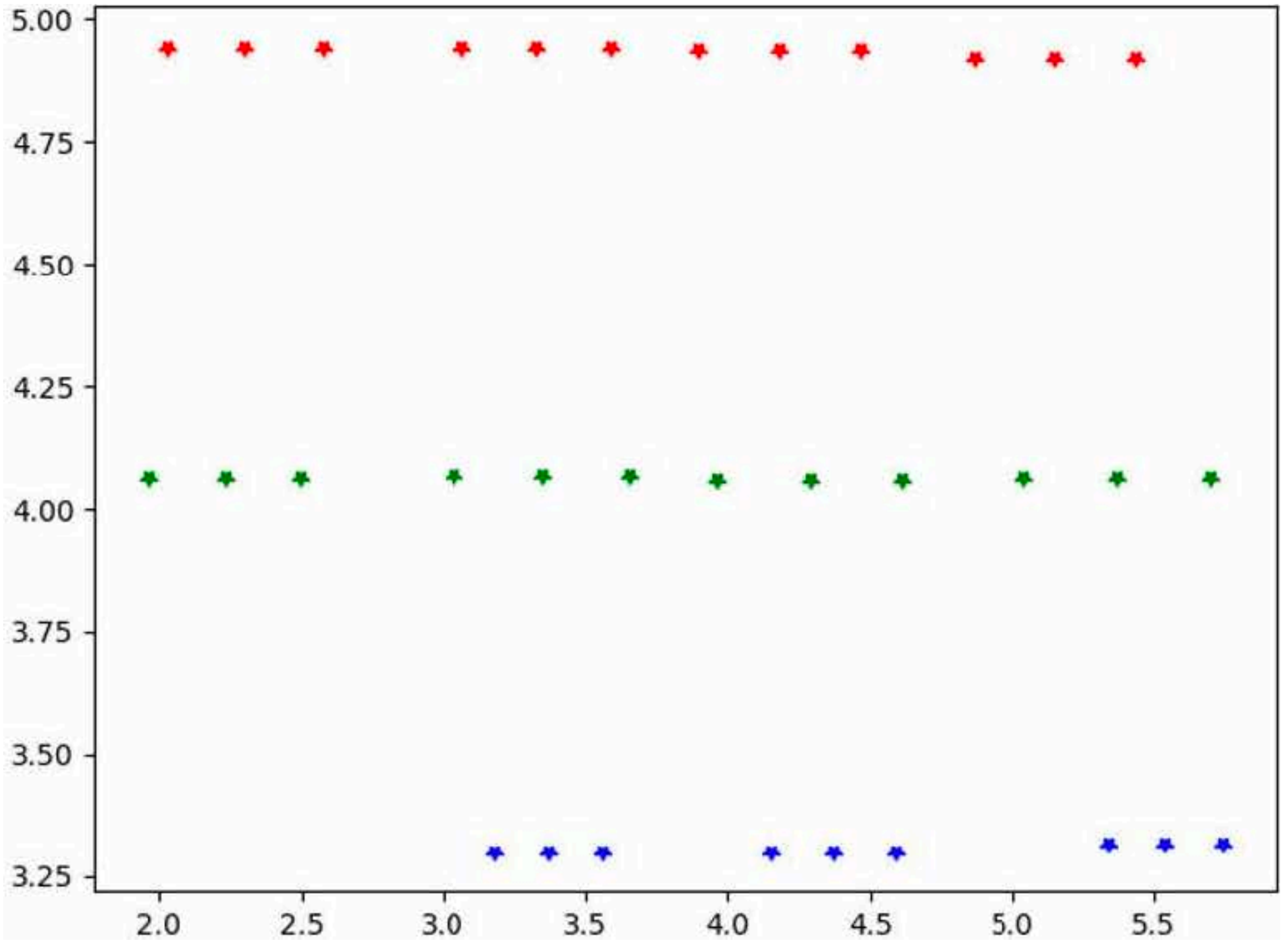
Results

Conclusions

Window detection + line fitting



[14]



Limitation: Restrictive rule-set