



# Apprentissage de modalités auxiliaires pour la localisation basée vision

Enhancing Visual-Based Localization by Learning Appearance of Paired Modalities

---

Nathan Piasco

26/06/2018

Congrès Reconnaissance des Formes, Image, Apprentissage et Perception 2018

# Introduction

# Visual Based Localization

Visual Based Localization (**VBL**) aims to recover the pose or position of a visual input query according to a known reference [Piasco et al., 2017].

$\rightarrow$ **?** $\rightarrow$

# Visual Based Localization: Indirect methods

VBL can be solved by retrieving the closest geo-referenced candidate in the database.

→


26.45 04.52


26.45 04.52


27.55 03.99


27.69 03.80


27.69 03.80


27.69 03.80


27.69 03.80

# Visual Based Localization: Indirect methods

VBL can be solved by retrieving the closest geo-referenced candidate in the database.

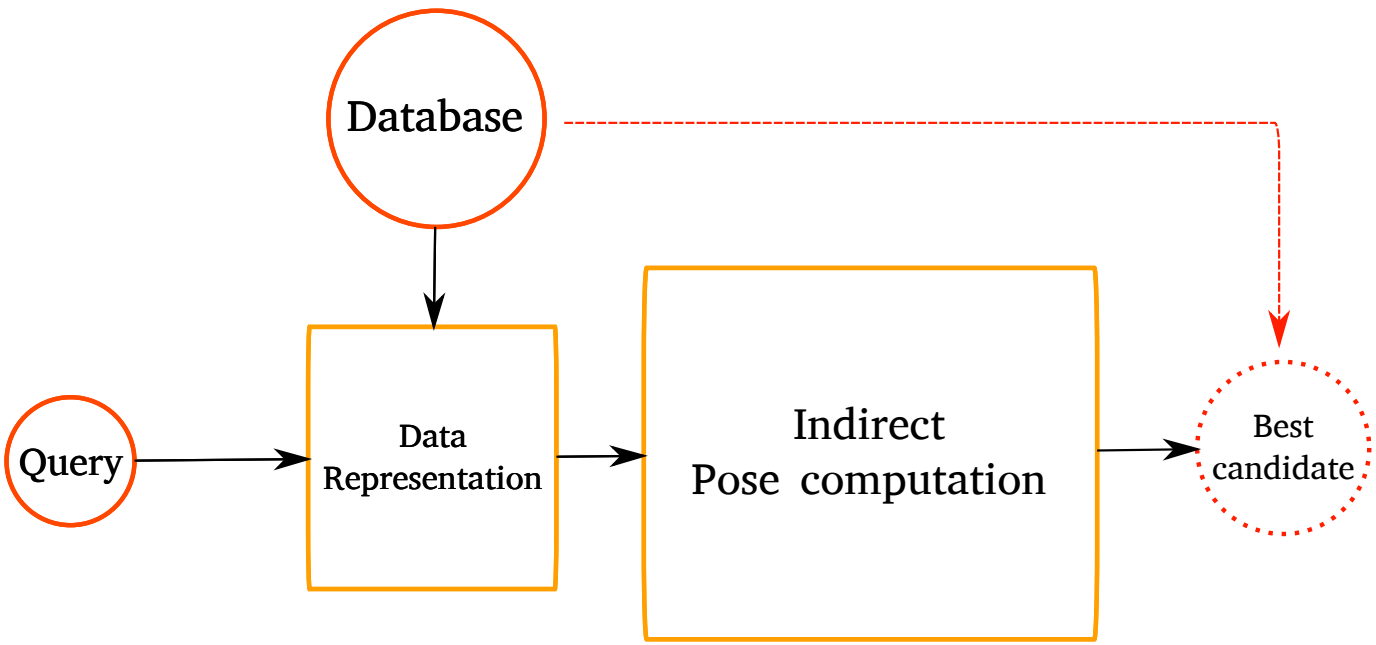
→

27.69 03.80

→

# Indirect methods: Pipeline

We propose a new data representation for solving indirect VBL.

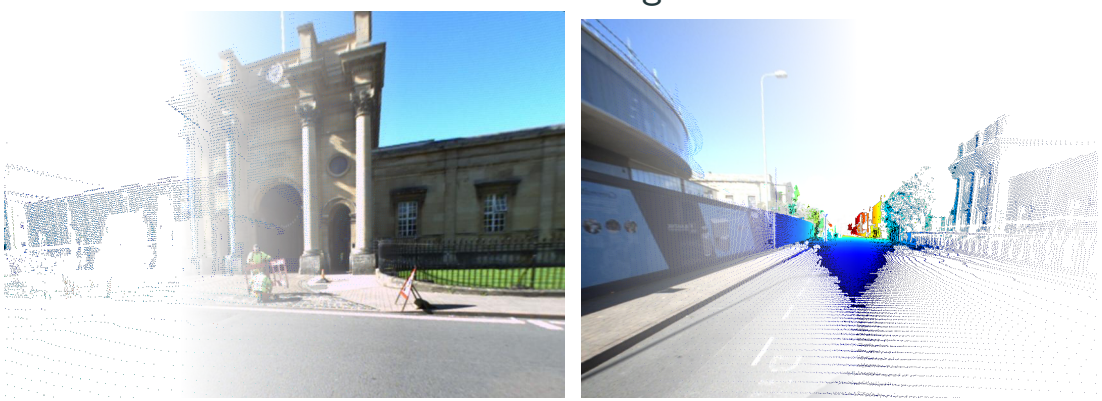


# Heterogeneous modalities

## Available data

| Training data type | Testing data type |
|---|---|
| **RGB + Depth** | RGB |

**Multi-modal** training dataset

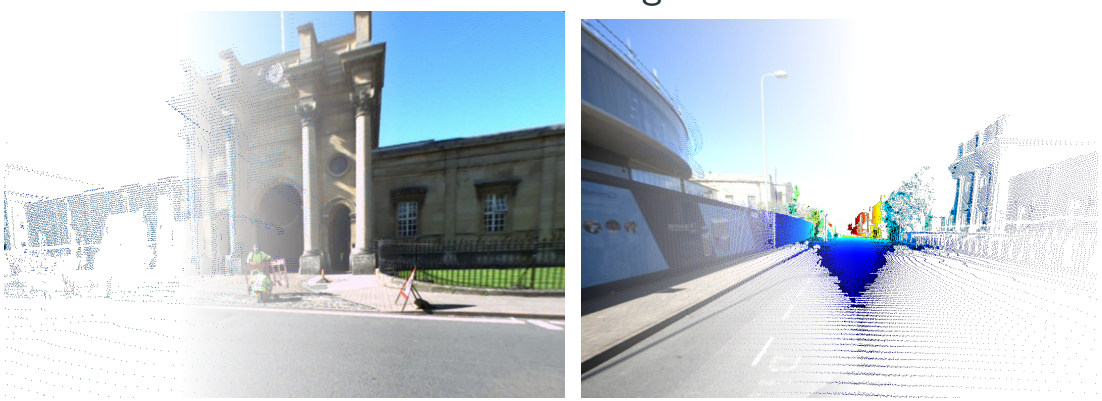


# Heterogeneous modalities

## Available data

| Training data type | Testing data type |
|---|---|
| RGB + Depth | **RGB** |

**Multi-modal** training dataset

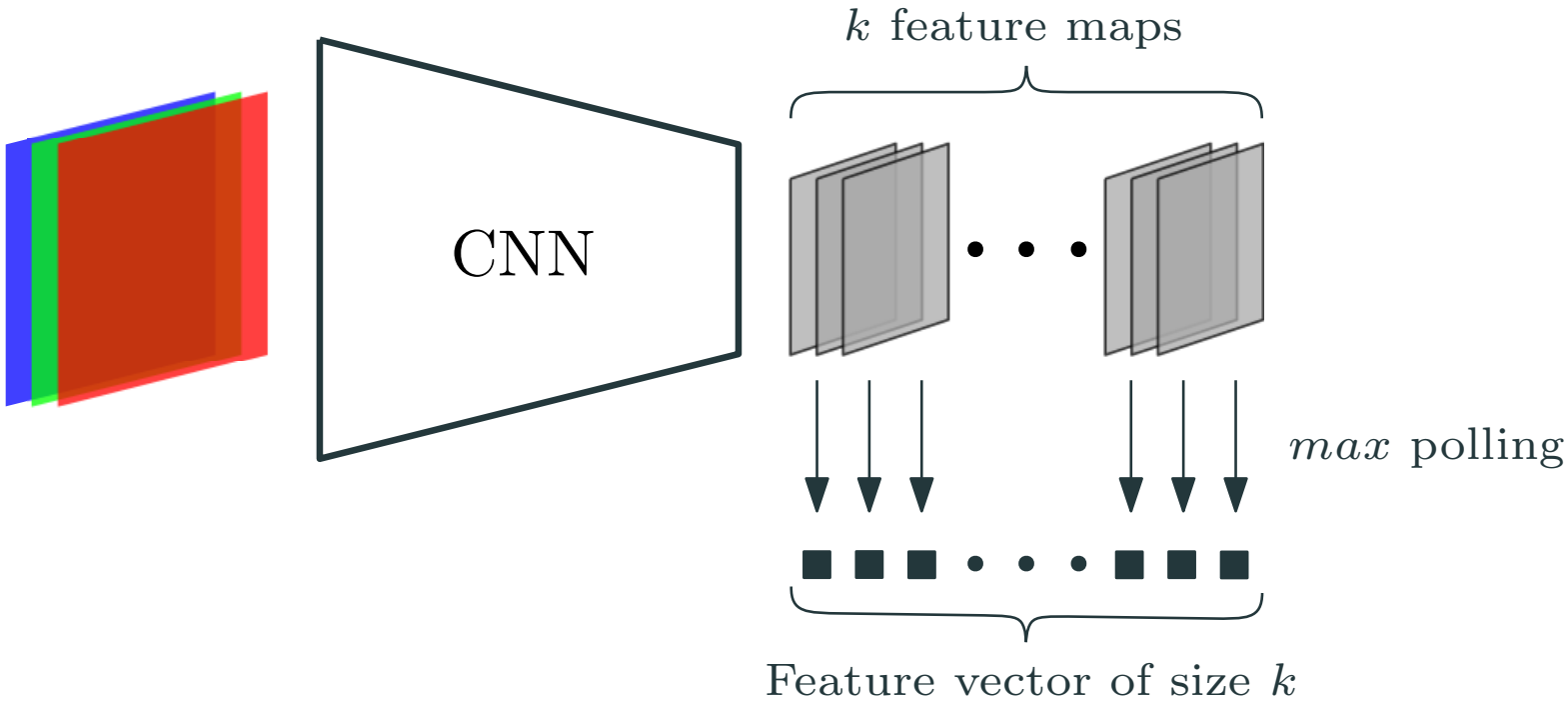


**Single modality** data at test time

# Related work

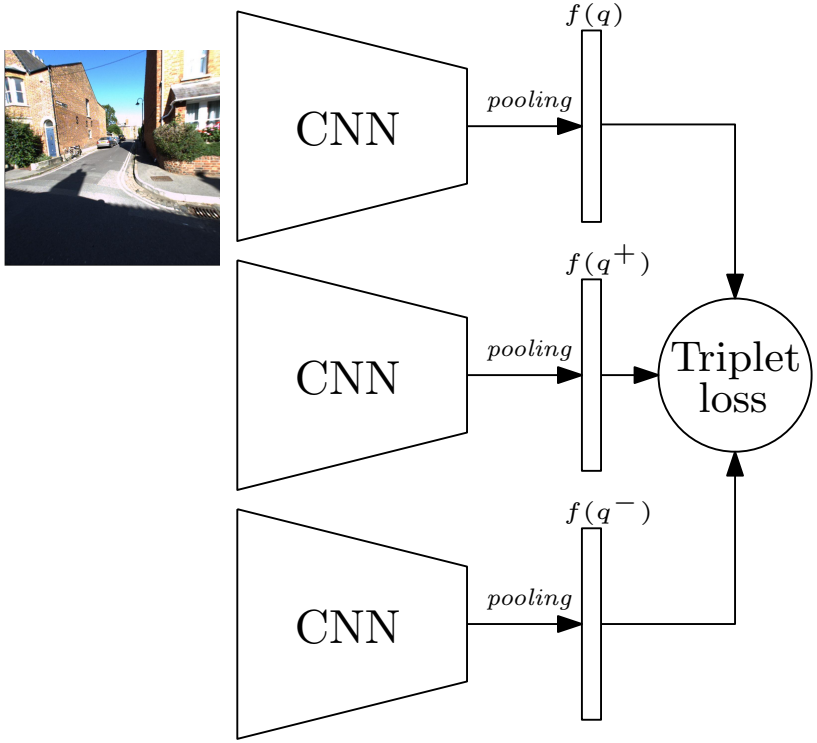
# Building a deep image descriptor

Fully connected part of a the network is dropped and pooling is done on the last convolutional response:




More complex aggregation methods exist:
NetVLAD [Arandjelović et al., 2017], RMAC...

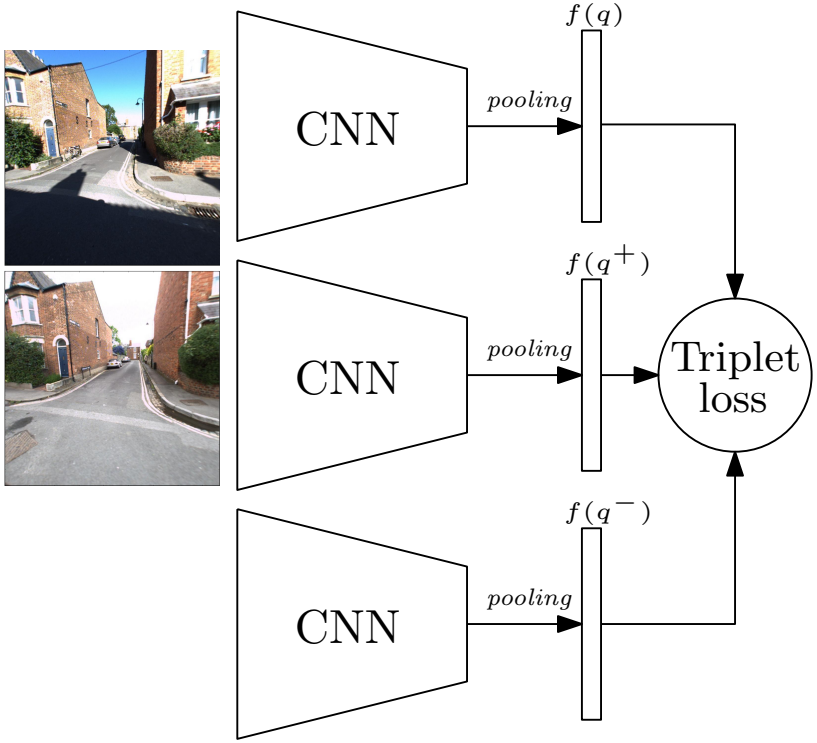
# Learning a deep image descriptor




$$Loss_{triplet} = max\left(\left\|f(q) - f(q^+)\right\|^2 - \left\|f(q) - f(q^-)\right\|^2 + \lambda, 0\right), \quad (1)$$

$$with \begin{cases} f(x) = \text{descriptor of image } x \\ \lambda = \text{triplet loss margin} \\ q = \text{query image} \\ q^+ = \text{positif example} \\ q^- = \text{negatif example} \end{cases}$$

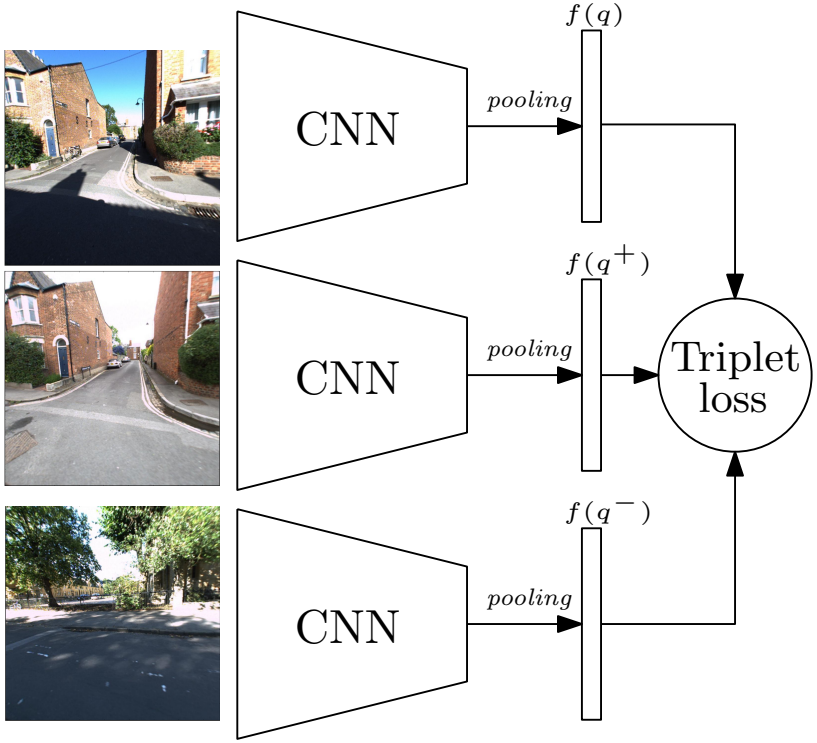
# Learning a deep image descriptor




$$Loss_{triplet} = max\left(\left\|f(q) - f(q^+)\right\|^2 - \left\|f(q) - f(q^-)\right\|^2 + \lambda, 0\right), \quad (1)$$

$$with \begin{cases} f(x) = \text{descriptor of image } x \\ \lambda = \text{triplet loss margin} \\ q = \text{query image} \\ q^+ = \text{positif example} \\ q^- = \text{negatif example} \end{cases}$$

# Learning a deep image descriptor

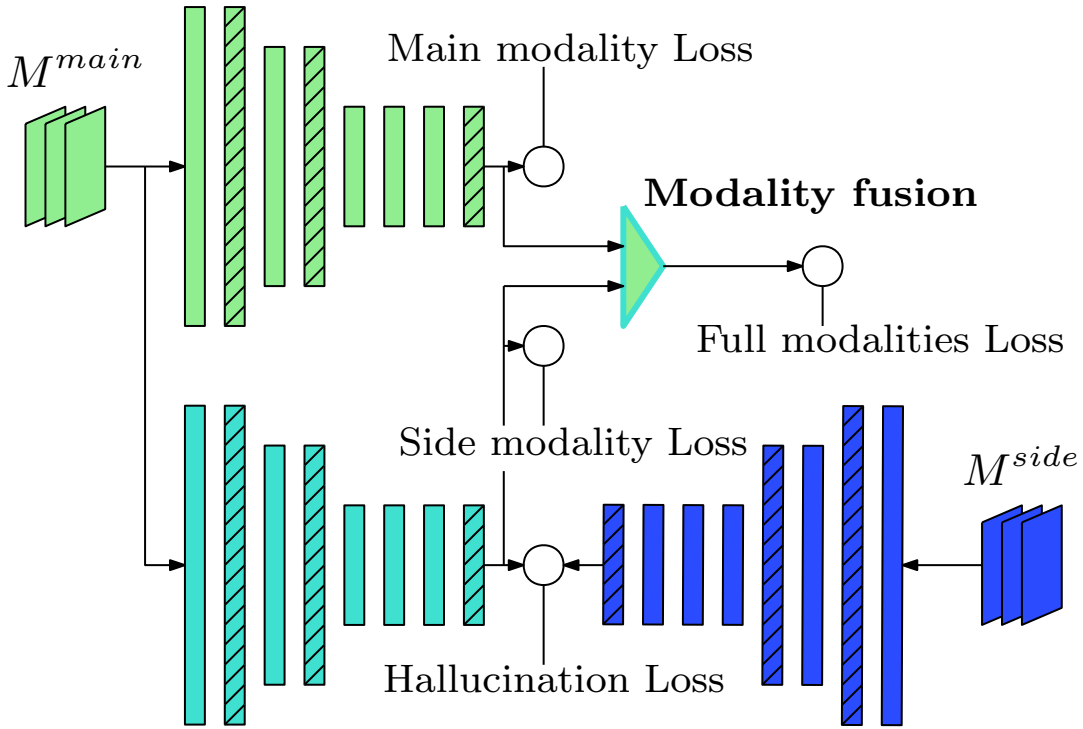



$$Loss_{triplet} = max\left(\left\|f(q) - f(q^+)\right\|^2 - \left\|f(q) - f(q^-)\right\|^2 + \lambda, 0\right), \quad (1)$$

$$with \begin{cases} f(x) = \text{descriptor of image } x \\ \lambda = \text{triplet loss margin} \\ q = \text{query image} \\ q^+ = \text{positif example} \\ q^- = \text{negatif example} \end{cases}$$

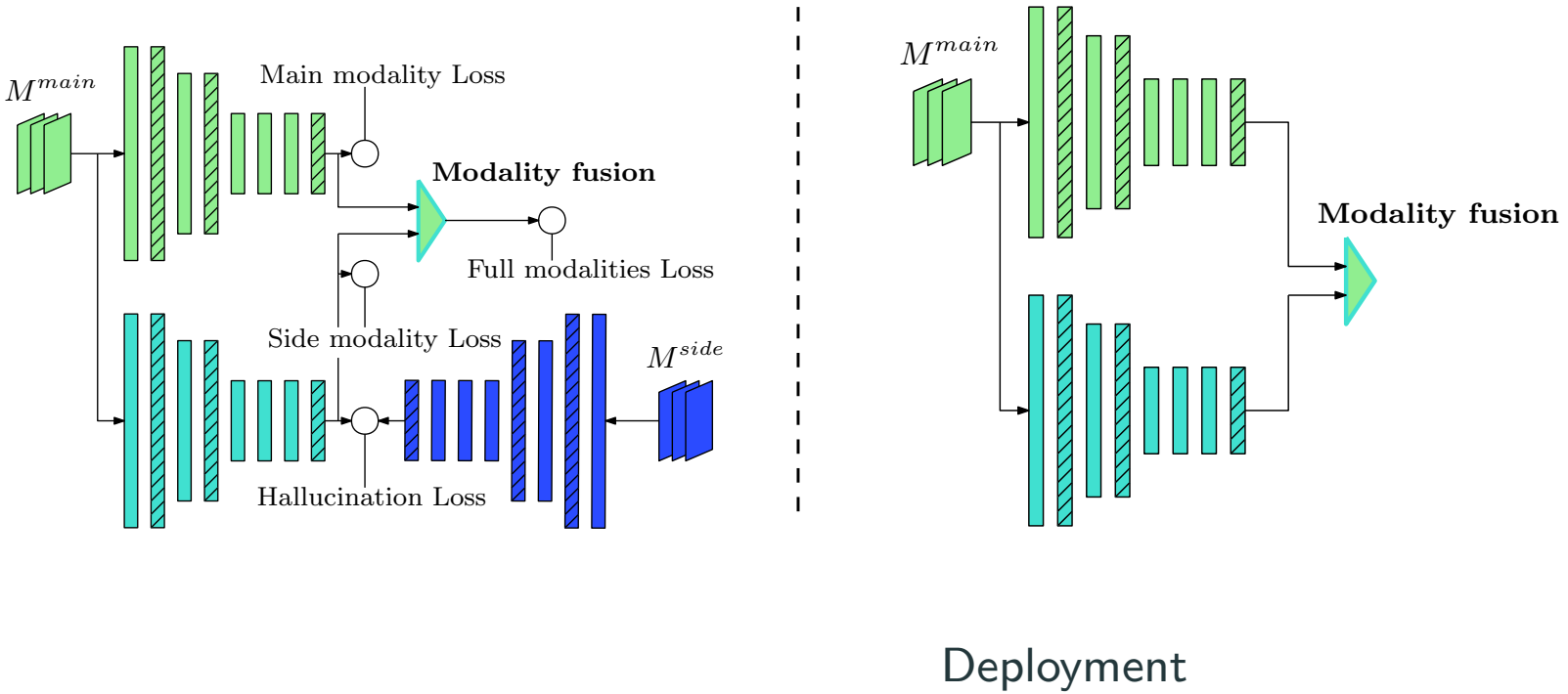
# Learning with side modality

Hallucination architecture from [Hoffman et al., 2016], **never been applied to image description and VBL**.



# Learning with side modality

Hallucination architecture from [Hoffman et al., 2016], **never been applied to image description and VBL**.



# Learning side information with modality transfer

# Intuition

Current encoder-decoder network architecture currently outperform all other methods for **modality transfer**.

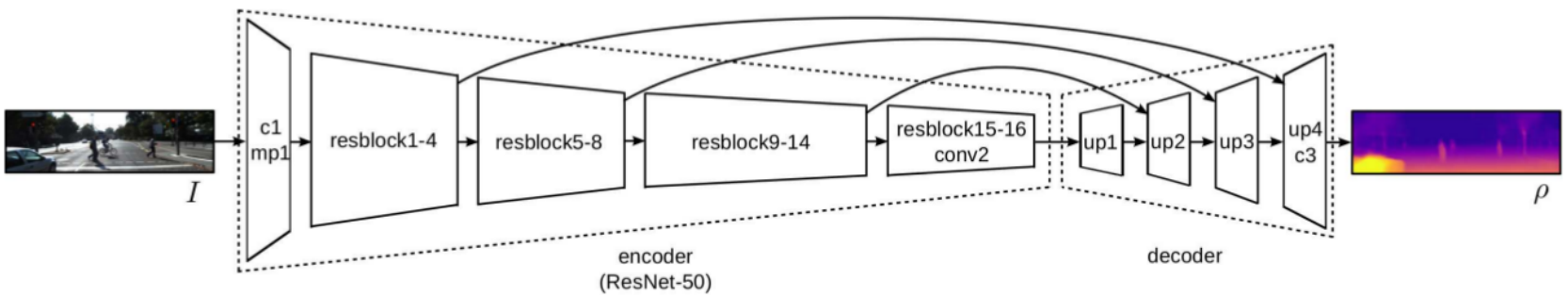
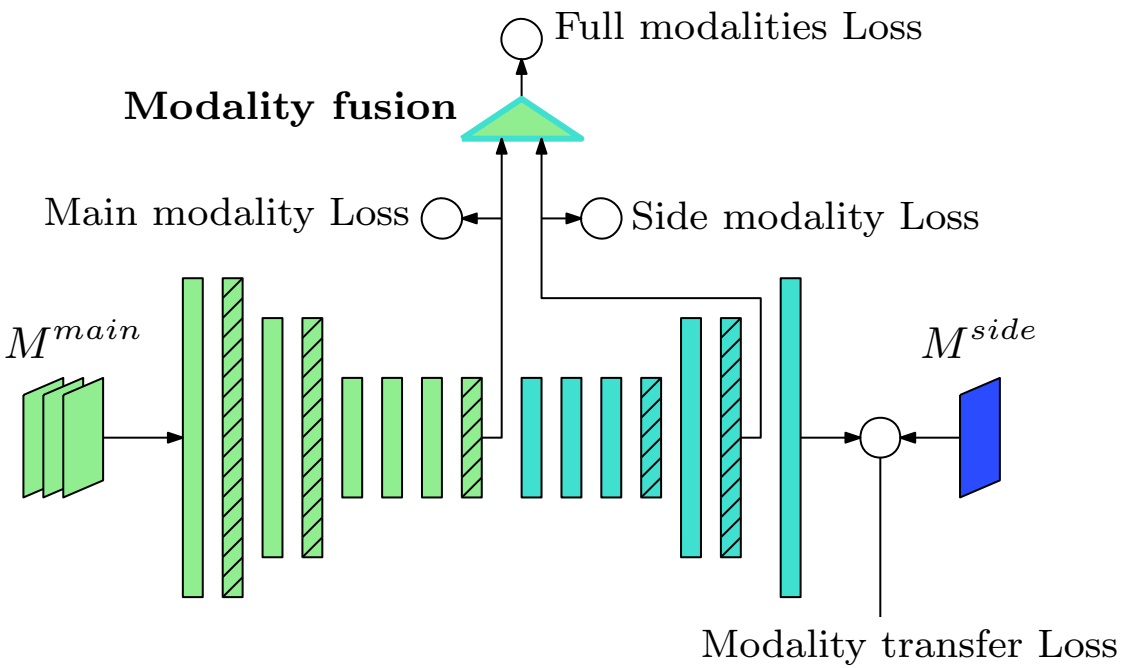



Illustration from [Kuznietsov et al., 2017]

# Proposed architecture

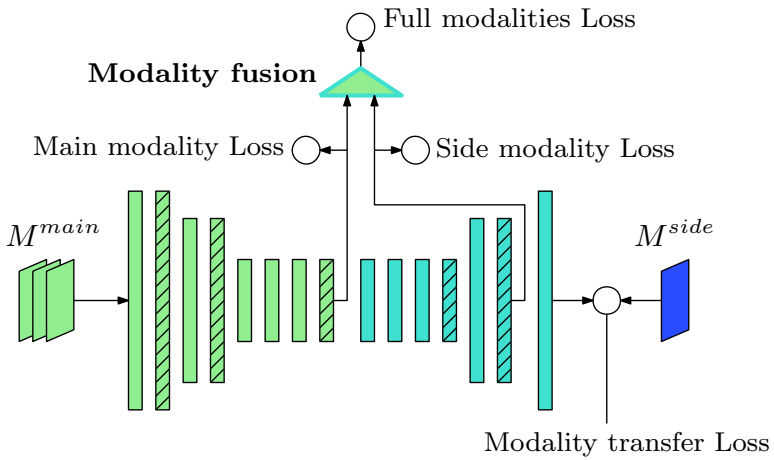
The proposed architecture inspired by encoder-decoder networks:

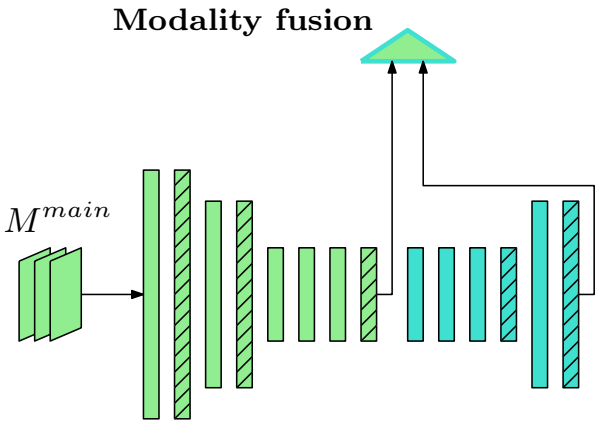



Training

# Proposed architecture

The proposed architecture inspired by encoder-decoder networks:




Deployment

# Optimization

$$Loss_{transfer} = \left\| \widetilde{M}(M^{main}) - M^{side} \right\|_1, \tag{2}$$

where $\widetilde{M}(x)$ denotes the output of the decoder part of the network regarding input $x$.

# Optimization

$$Loss_{transfer} = \left\| \widetilde{M}(M^{main}) - M^{side} \right\|_1, \tag{2}$$

where $\widetilde{M}(x)$ denotes the output of the decoder part of the network regarding input $x$.

Final loss:

$$Loss = Loss^{main}_{triplet} + Loss^{side}_{triplet} * \sigma_{side} + Loss^{full}_{triplet} * \sigma_{full} + Loss_{transfer} * \sigma_{transfer}. \tag{3}$$

# Optimization

$$Loss_{transfer} = \left\| \widetilde{M}(M^{main}) - M^{side} \right\|_1, \tag{2}$$

where $\widetilde{M}(x)$ denotes the output of the decoder part of the network regarding input $x$.

Final loss:

$$Loss = Loss^{main}_{triplet} + Loss^{side}_{triplet} * \sigma_{side} + Loss^{full}_{triplet} * \sigma_{full} + Loss_{transfer} * \sigma_{transfer}. \tag{3}$$

Diversification loss:

$$Loss_{div} = max\left(Loss^{full}_{triplet} - Loss^{main}_{triplet} + \lambda_{div}, 0\right), \tag{4}$$

where $\lambda_{div}$ is a scalar value that acts as a margin to ensure $Loss^{full}_{triplet}$ is always smaller than $Loss^{main}_{triplet}$.

# Advantages over hallucination

Advantages of our bottom-up transfer approach (**BU-TF**) over hallucination network are threefold:

- No need of pretraining on side modality

# Advantages over hallucination

Advantages of our bottom-up transfer approach (**BU-TF**) over hallucination network are threefold:

- No need of pretraining on side modality
- Method by nature lighter: $29k$ parameters vs. $41k$ parameters for networks built upon Alexnet architecture

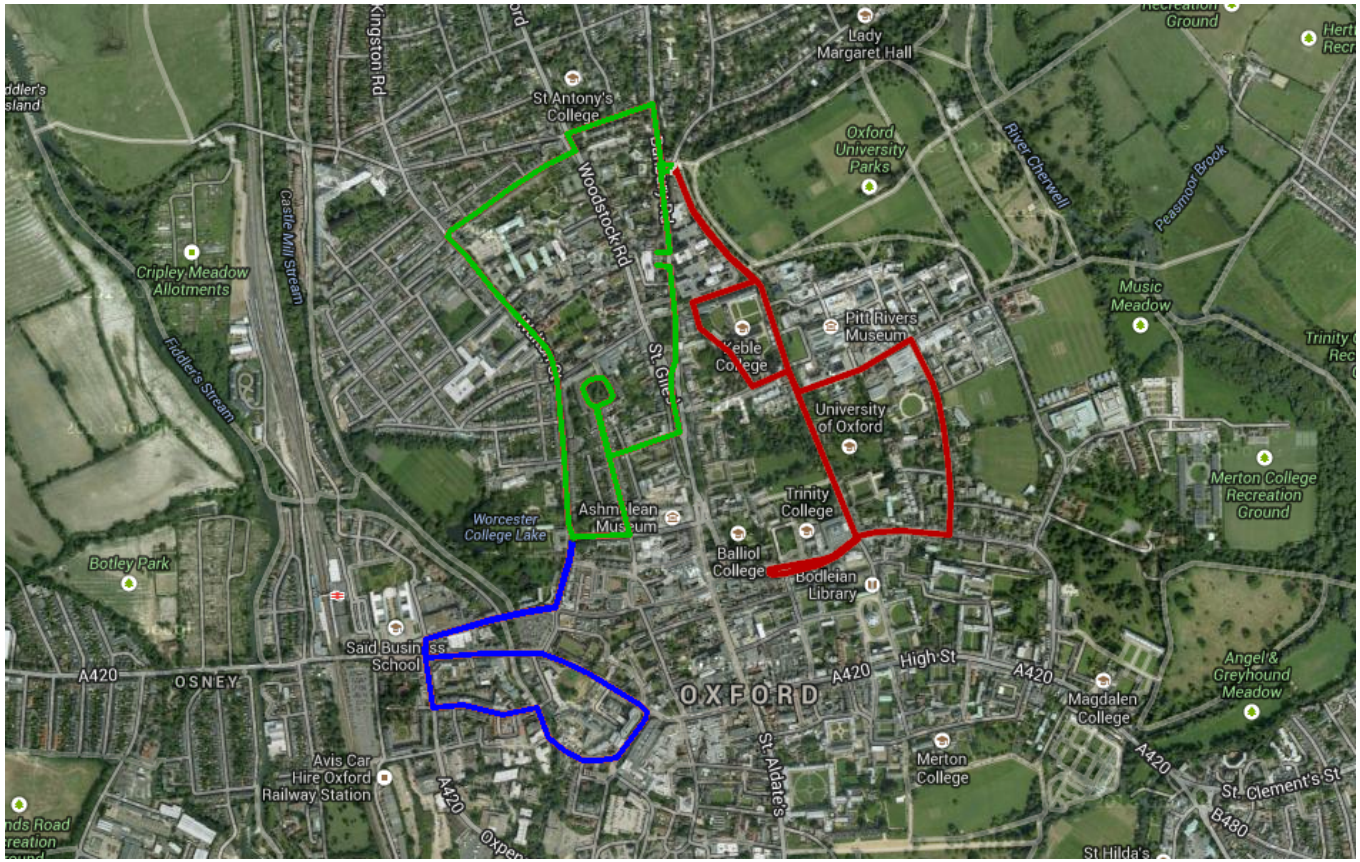
# Advantages over hallucination

Advantages of our bottom-up transfer approach (**BU-TF**) over hallucination network are threefold:

- No need of pretraining on side modality
- Method by nature lighter: $29k$ parameters vs. $41k$ parameters for networks built upon Alexnet architecture
- No need to transform modality into 3-channels data

# Experiments

# Robotcar dataset



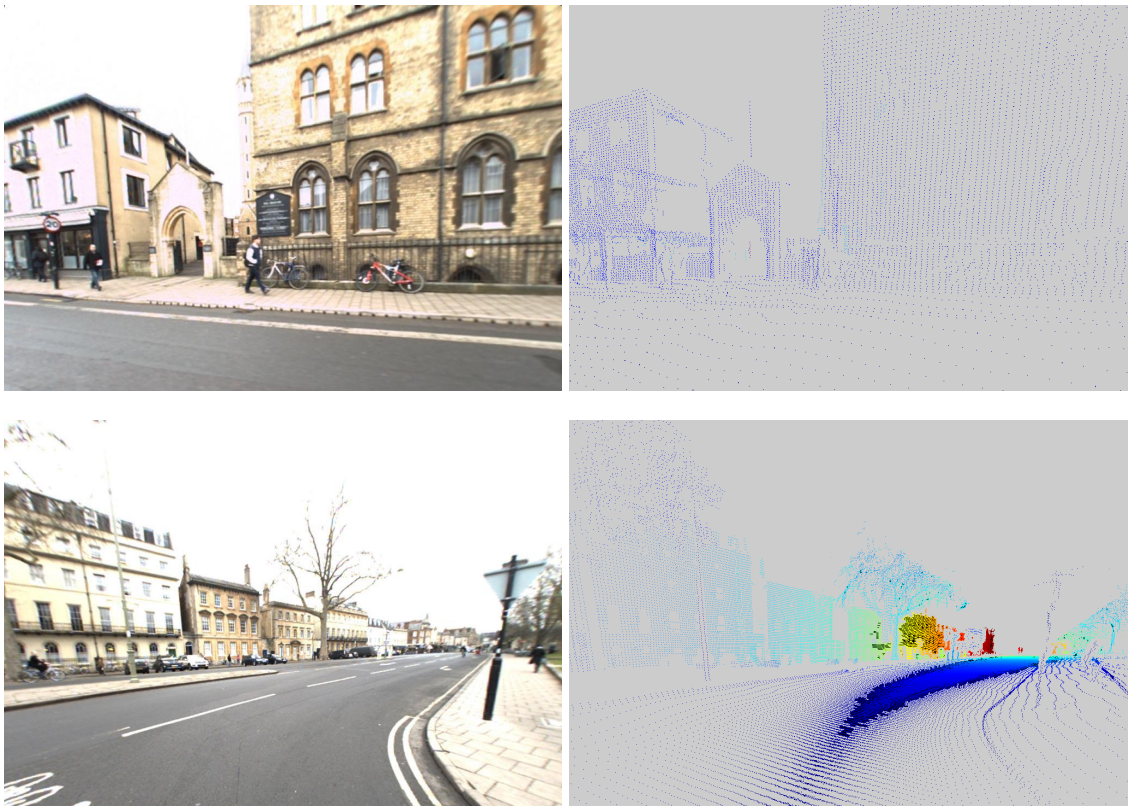

Dataset training (green), validation (blue) and testing (red) areas.

# Robotcar dataset

Examples of queries with corresponding dataset candidates of the testing set.

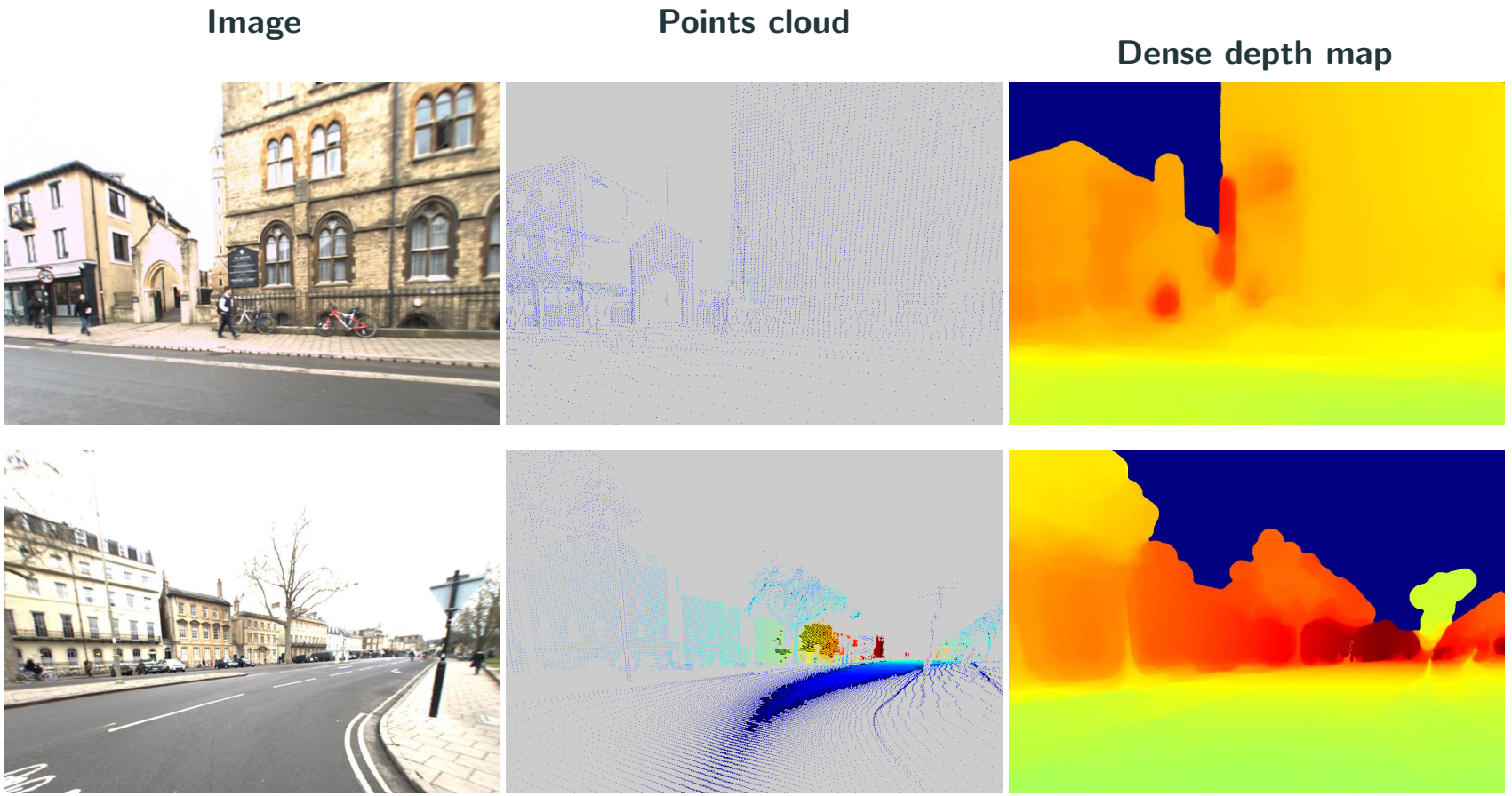
# Building dense modality map

**Image** **Points cloud**



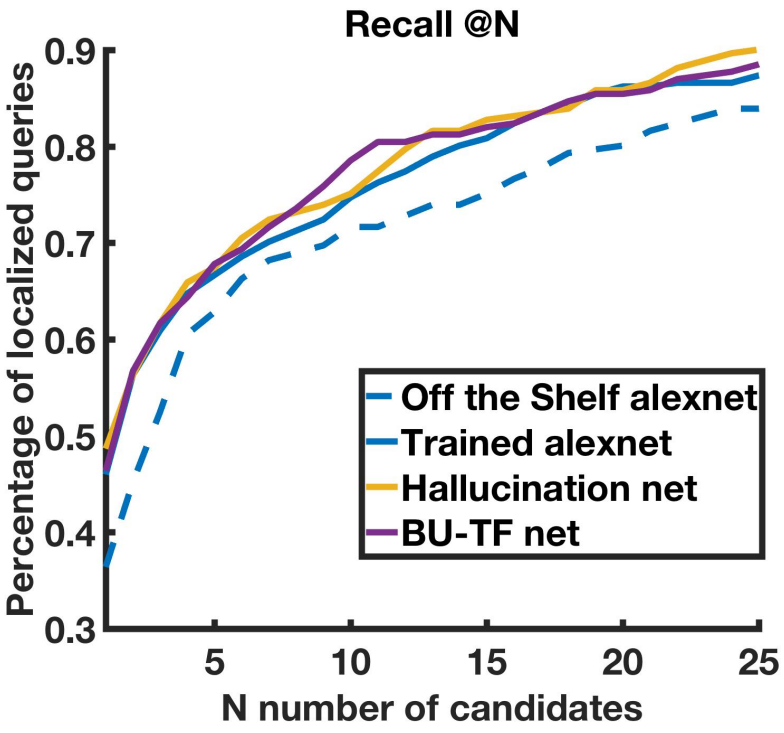
We use the algorithm proposed in [Bevilacqua et al., 2017] to create a dense modality map from an image and the associated point cloud.

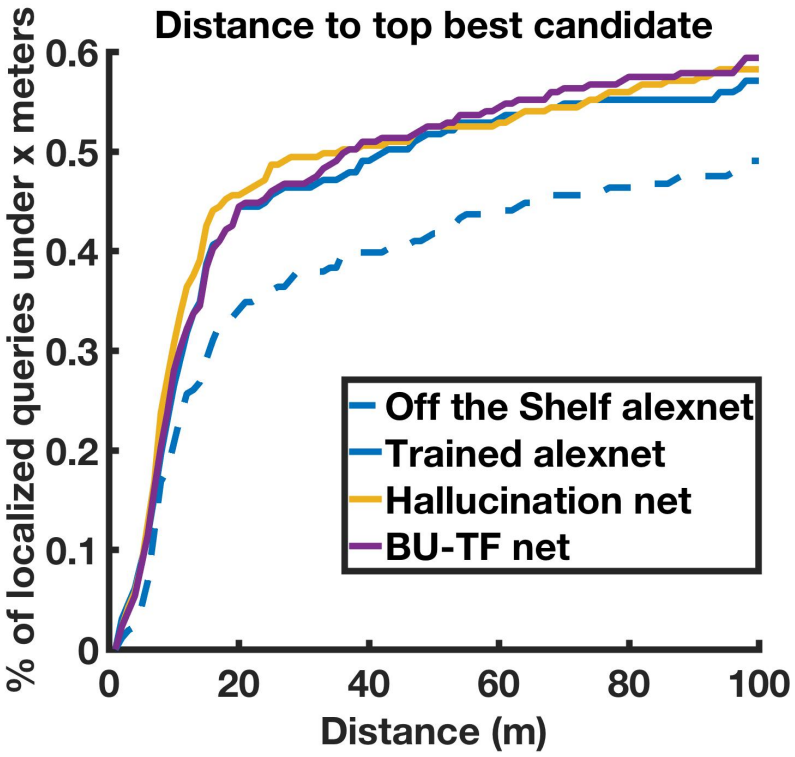
# Building dense modality map




We use the algorithm proposed in [Bevilacqua et al., 2017] to create a dense modality map from an image and the associated point cloud.

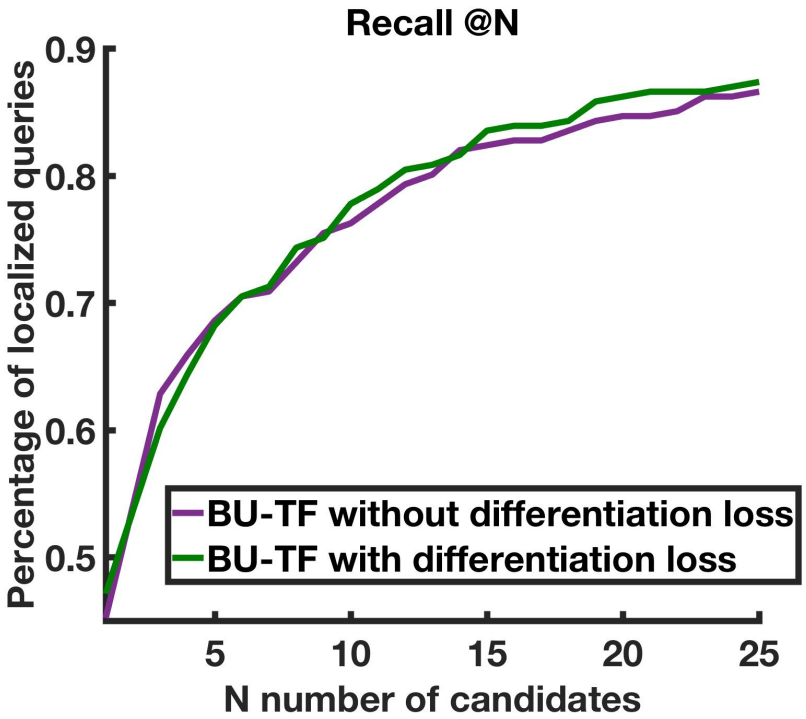
# Results

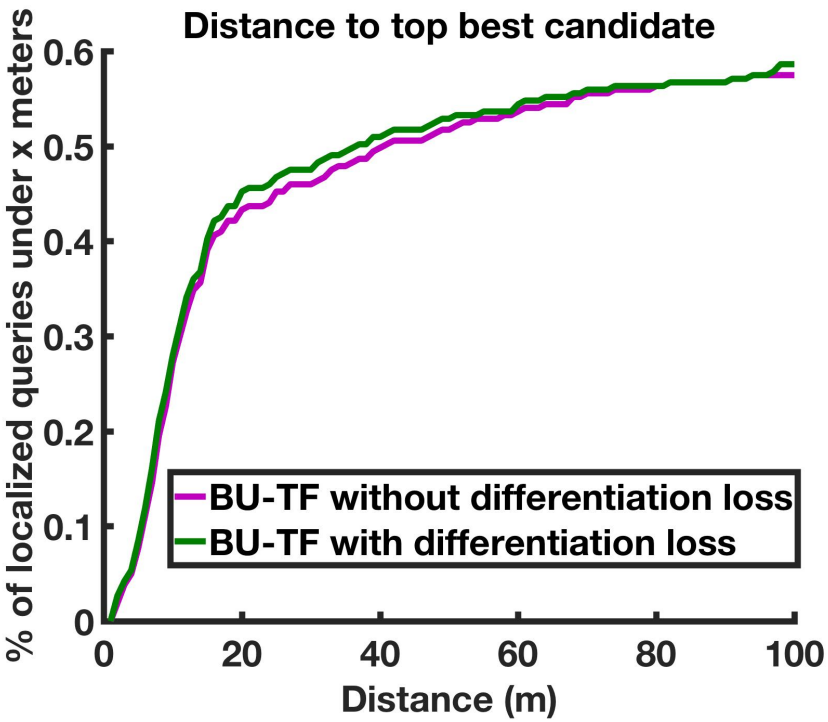



Off-the-shelf: network only trained on ImageNet, no fine-tuning for this specific task and on these specific data.

# Results - Diversification loss



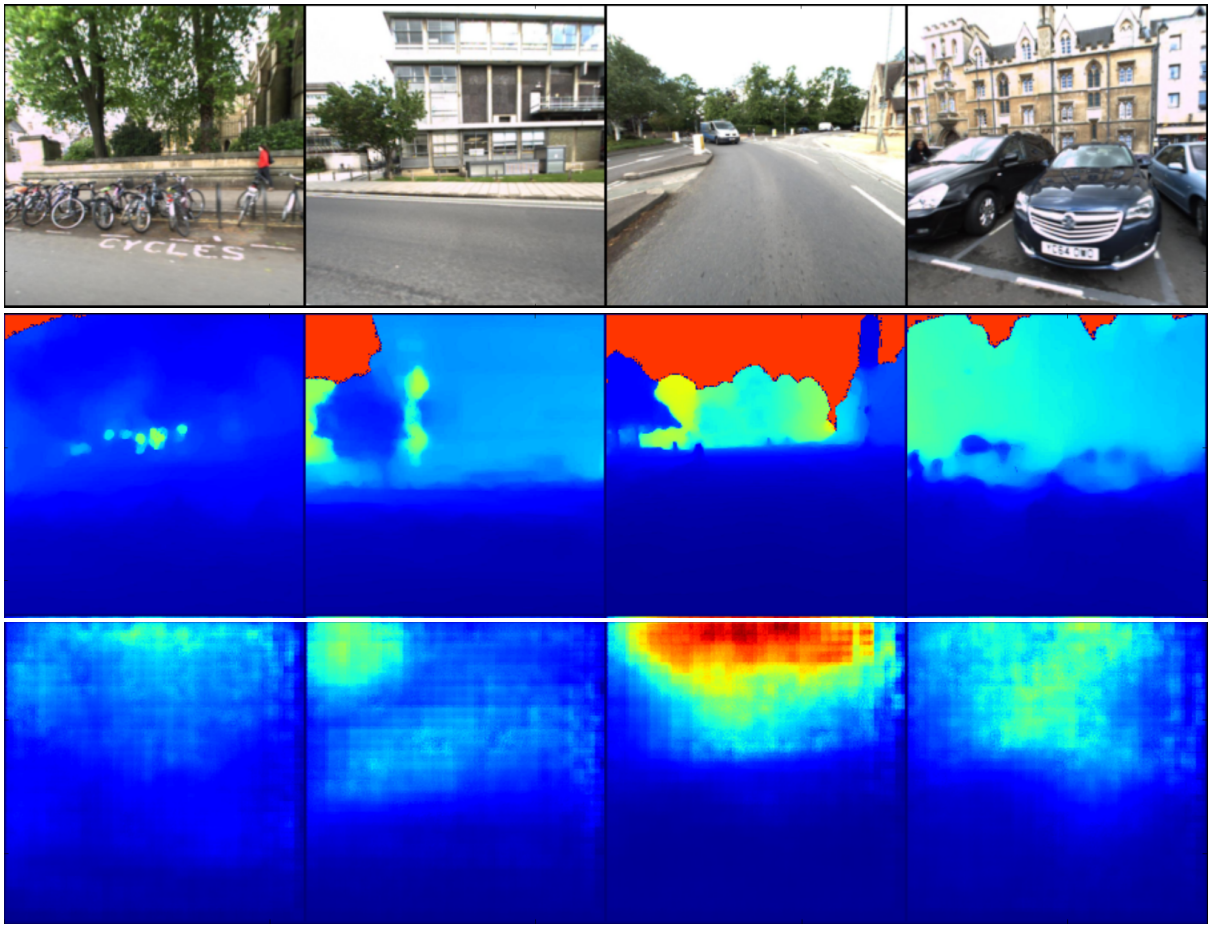



# Results - Visual inspection

Main modality

Side modality

Reconstructed side modality



# Conclusion

# Conclusion

New method for learning with modality side modality have been presented. BU-TF is more efficient than hallucination as it needs less training time have less parameters while producing comparable results.

# Conclusion

New method for learning with modality side modality have been presented. BU-TF is more efficient than hallucination as it needs less training time have less parameters while producing comparable results.

*Future work* – The presented method has to be tested:

- on other modalities

# Conclusion

New method for learning with modality side modality have been presented. BU-TF is more efficient than hallucination as it needs less training time have less parameters while producing comparable results.

*Future work* – The presented method has to be tested:

- on other modalities
- with other aggregation scheme

# Conclusion

New method for learning with modality side modality have been presented. BU-TF is more efficient than hallucination as it needs less training time have less parameters while producing comparable results.

*Future work* – The presented method has to be tested:

- on other modalities
- with other aggregation scheme
- on other visual localisation tasks (*e.g.* pose regression)

**Thanks for your attention**

*Question time*

# References I



Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. (2017).
**NetVLAD: CNN architecture for weakly supervised place recognition.**
*IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 5297–5307.


Bevilacqua, M., Aujol, J. F., Biasutti, P., Brédif, M., and Bugeau, A. (2017).
**Joint inpainting of depth and reflectance with visibility estimation.**
*ISPRS Journal of Photogrammetry and Remote Sensing*, 125:16–32.


Hoffman, J., Gupta, S., and Darrell, T. (2016).
**Learning with Side Information through Modality Hallucination.**
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 826–834.

# References II


Kuznietsov, Y., Stückler, J., and Leibe, B. (2017).
**Semi-Supervised Deep Learning for Monocular Depth Map Prediction.**
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Piasco, N., Sidibé, D., Demonceaux, C., and Gouet-Brunet, V. (2017).
**A survey on Visual-Based Localization: On the benefit of heterogeneous data.**
*Pattern Recognition*, 74:90–109.