



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

Related work

Our proposal: Learning side information with modality transfer

Experiments

Conclusion

Introduction

Visual Based Localization

Visual Based Localization (**VBL**) aims to recover the pose or position of a visual input query according to a known visual reference.

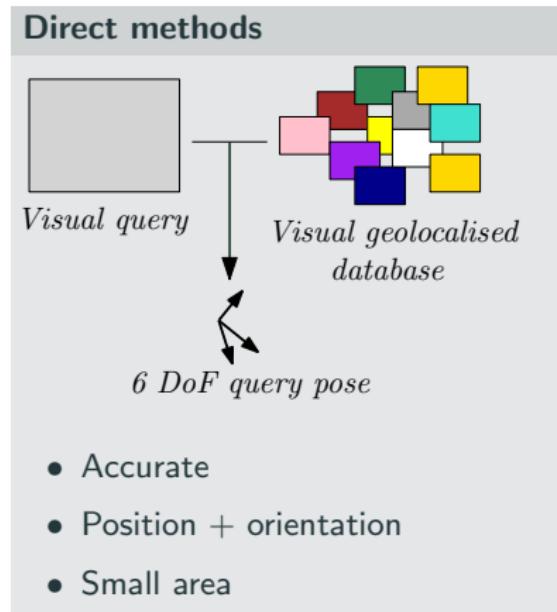


→ ? →



Direct vs Indirect methods

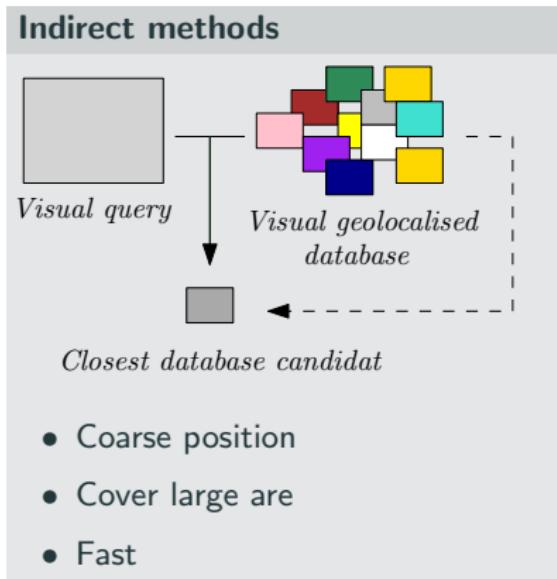
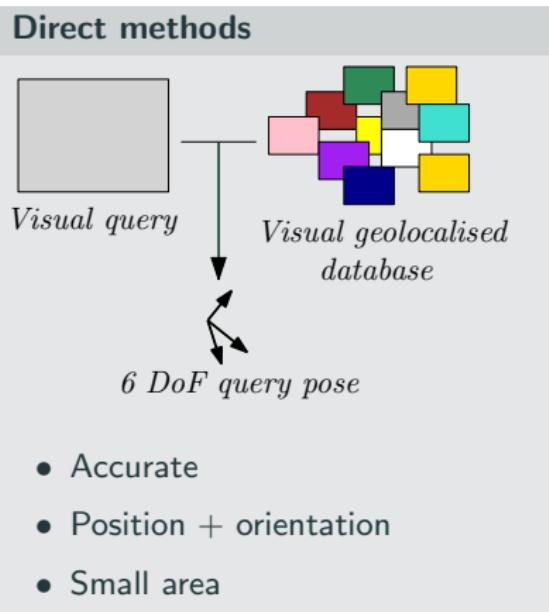
Visual Based Localization (**VBL**) methods can be divided in two main categories [Piasco et al., 2017]:



- Accurate
- Position + orientation
- Small area

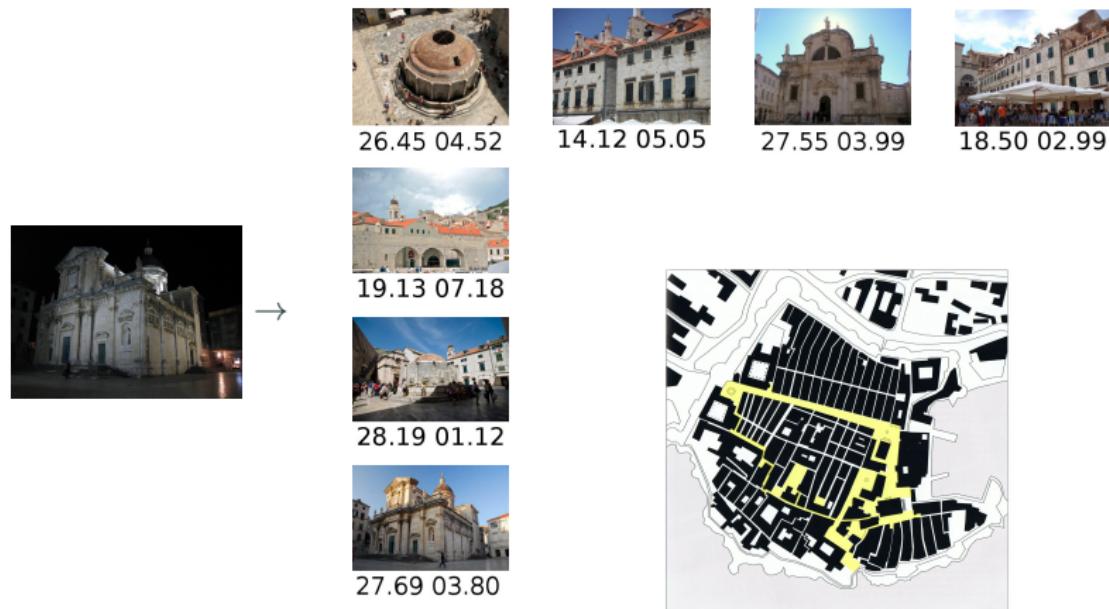
Direct vs Indirect methods

Visual Based Localization (**VBL**) methods can be divided in two main categories [Piasco et al., 2017]:



Visual Based Localization: Indirect methods

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



Visual Based Localization: Indirect methods

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

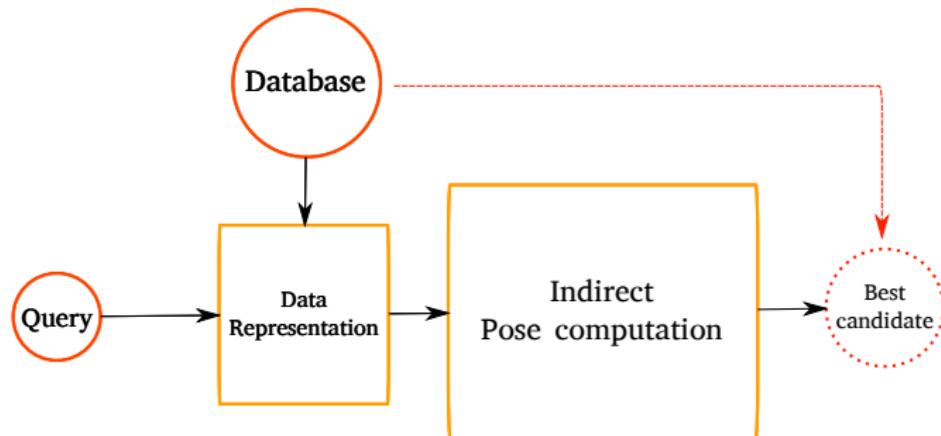


27.69 03.80



Indirect methods: Pipeline

We propose a new data representation for solving indirect VBL.



Various modalities

In this work, we tackle the problem of VBL with **not only** images.

Modalities that could be used

- Image (main modality)



Introduction

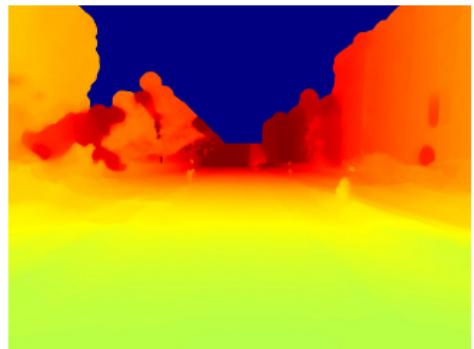


Various modalities

In this work, we tackle the problem of VBL with **not only** images.

Modalities that could be used

- **Image** (main modality)
- Depth map

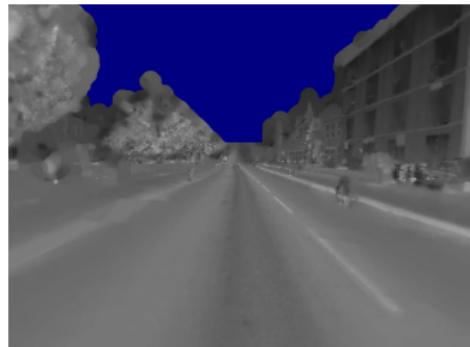


Various modalities

In this work, we tackle the problem of VBL with **not only** images.

Modalities that could be used

- **Image** (main modality)
- Depth map
- Reflectance



INSTITUT NATIONAL
DE L'INFORMATIQUE
DES TELECOMS
ET DES SYSTEMES

Introduction



UNIVERSITE
BOURGOGNE FRANCHE-COMTE



Various modalities

In this work, we tackle the problem of VBL with **not only** images.

Modalities that could be used

- **Image** (main modality)
- Depth map
- Reflectance
- Semantic information
- ...



Heterogeneous modalities

Available data

Training data type	Testing data type
RGB + Depth	RGB

Multi-modal training dataset



Introduction

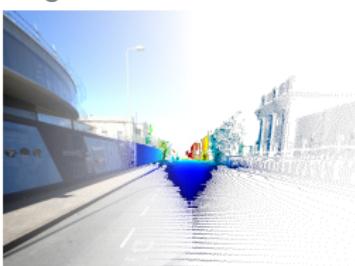


Heterogeneous modalities

Available data

Training data type	Testing data type
RGB + Depth	RGB

Multi-modal training dataset



Single modality data at test



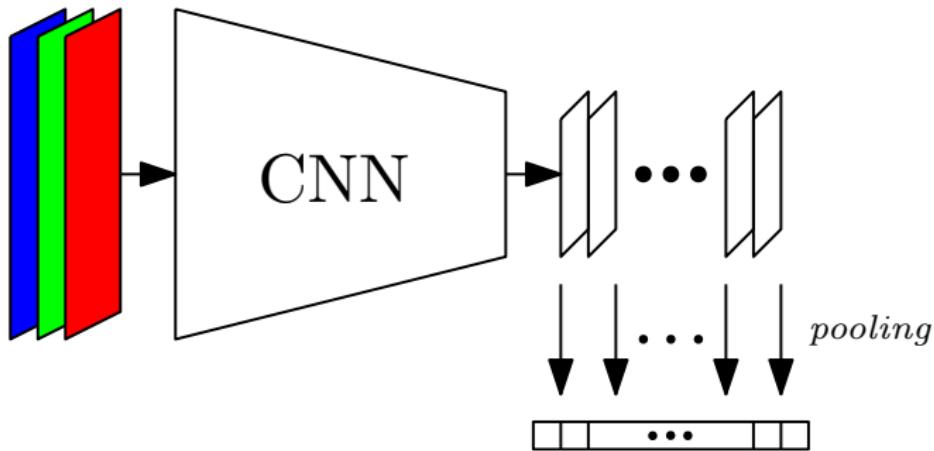
Introduction



Related work

Building a deep image descriptor

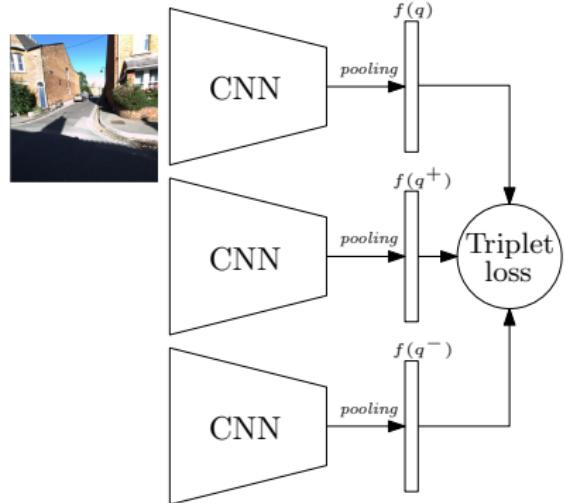
Fully connected part of a the network is dropped and pooling is done on the last convolutional response:



More complex aggregation methods exist:

NetVLAD [Arandjelović et al., 2017], RMAC [Radenović et al., 2016]...

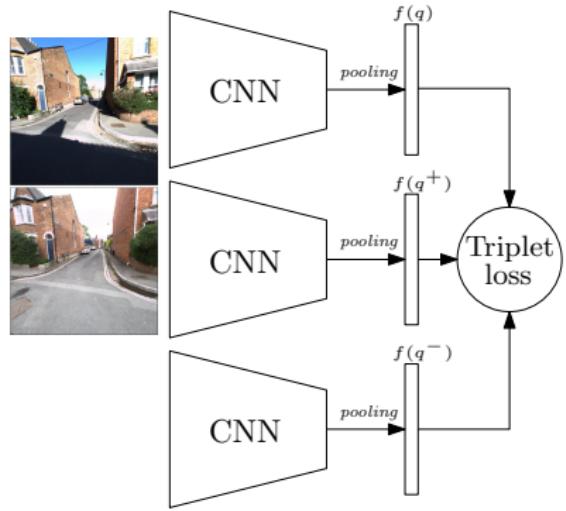
Learning a deep image descriptor



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

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

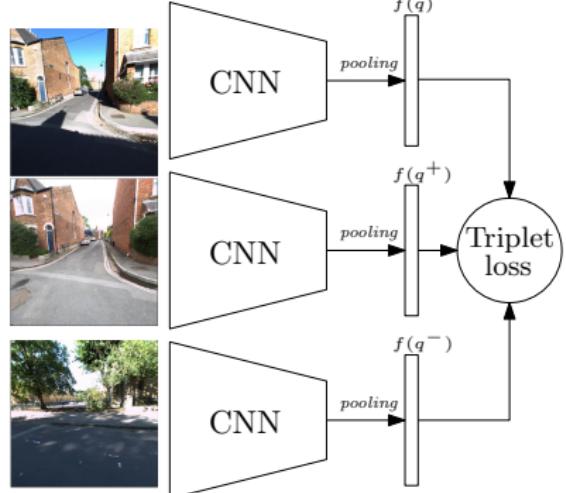
Learning a deep image descriptor



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

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

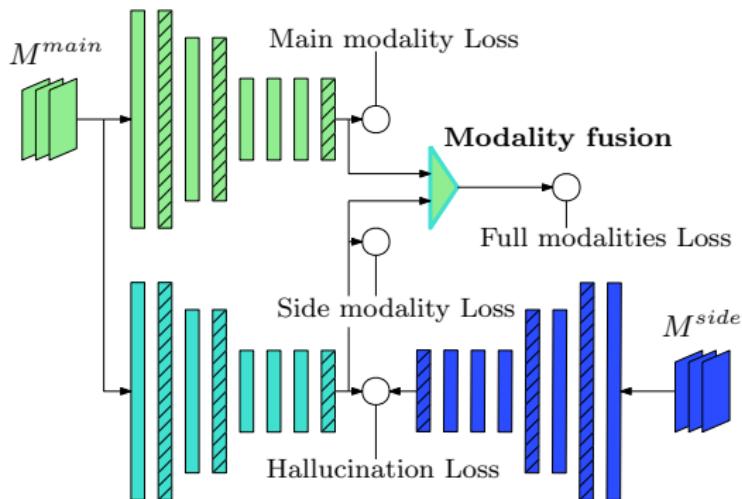


$$Loss_{triplet} = \max \left(\|f(q) - f(q^+)\|^2 - \|f(q) - f(q^-)\|^2 + \lambda, 0 \right), \quad (1)$$

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

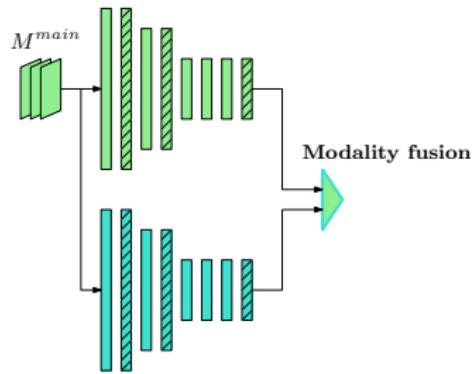
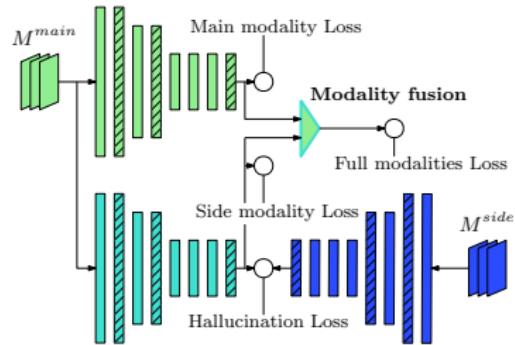
Hallucination architecture from [Hoffman et al., 2016] for objects classification, never applied to image description and VBL.



Related work

Learning with side modality

Hallucination architecture from [Hoffman et al., 2016] for objects classification, never applied to image description and VBL.



Deployment

Our proposal: Learning side information with modality transfer

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

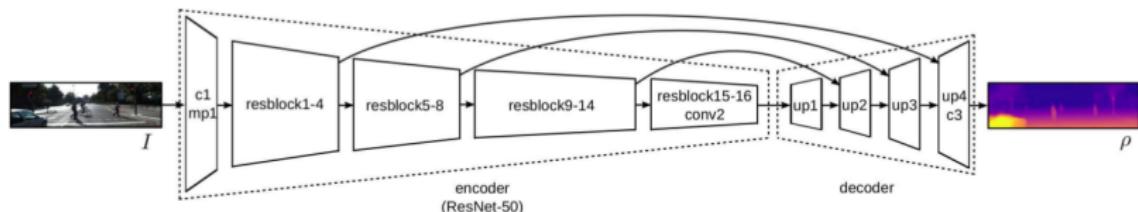
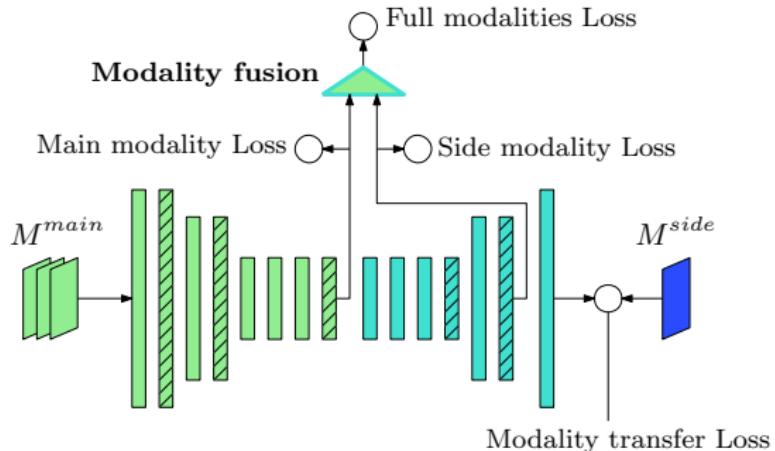


Illustration from [Kuznetsov et al., 2017]

Proposed architecture

The proposed architecture is inspired by encoder-decoder networks:



Training

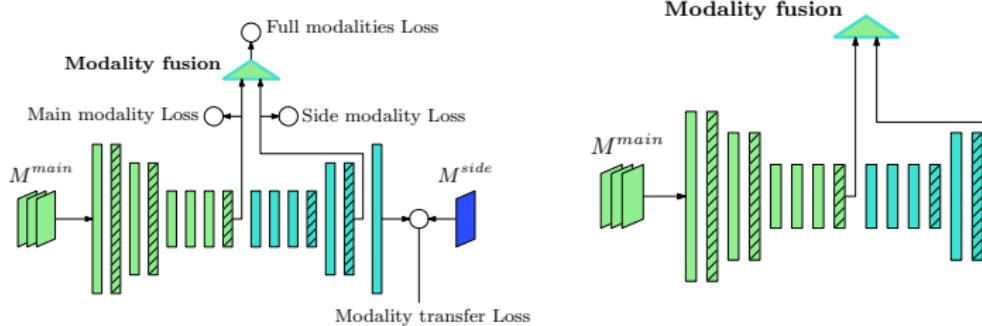


Our proposal: Learning side information with modality transfer



Proposed architecture

The proposed architecture is inspired by encoder-decoder networks:



Deployment

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

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

Optimization

$$\text{Loss}_{\text{transfer}} = \left\| \tilde{M}(M^{\text{main}}) - M^{\text{side}} \right\|_1, \quad (2)$$

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

Final loss:

$$\begin{aligned} \text{Loss} = & \text{Loss}_{\text{triplet}}^{\text{main}} + \text{Loss}_{\text{triplet}}^{\text{side}} * \sigma_{\text{side}} \\ & + \text{Loss}_{\text{triplet}}^{\text{full}} * \sigma_{\text{full}} + \text{Loss}_{\text{transfer}} * \sigma_{\text{transfer}}. \end{aligned} \quad (3)$$

Optimization

$$\text{Loss}_{\text{transfer}} = \left\| \tilde{M}(M^{\text{main}}) - M^{\text{side}} \right\|_1, \quad (2)$$

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

Final loss:

$$\begin{aligned} \text{Loss} = & \text{Loss}_{\text{triplet}}^{\text{main}} + \text{Loss}_{\text{triplet}}^{\text{side}} * \sigma_{\text{side}} \\ & + \text{Loss}_{\text{triplet}}^{\text{full}} * \sigma_{\text{full}} + \text{Loss}_{\text{transfer}} * \sigma_{\text{transfer}}. \end{aligned} \quad (3)$$

Diversification loss:

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

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

Advantages over hallucination

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

- No need of pretraining on side modality



INSTITUT NATIONAL
DE L'INFORMATION
GÉOGRAPHIQUE
ET forestière

Our proposal: Learning side information with modality transfer



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



INSTITUT NATIONAL
DE L'INFORMATION
GÉOGRAPHIQUE
ET forestière

Our proposal: Learning side information with modality transfer



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



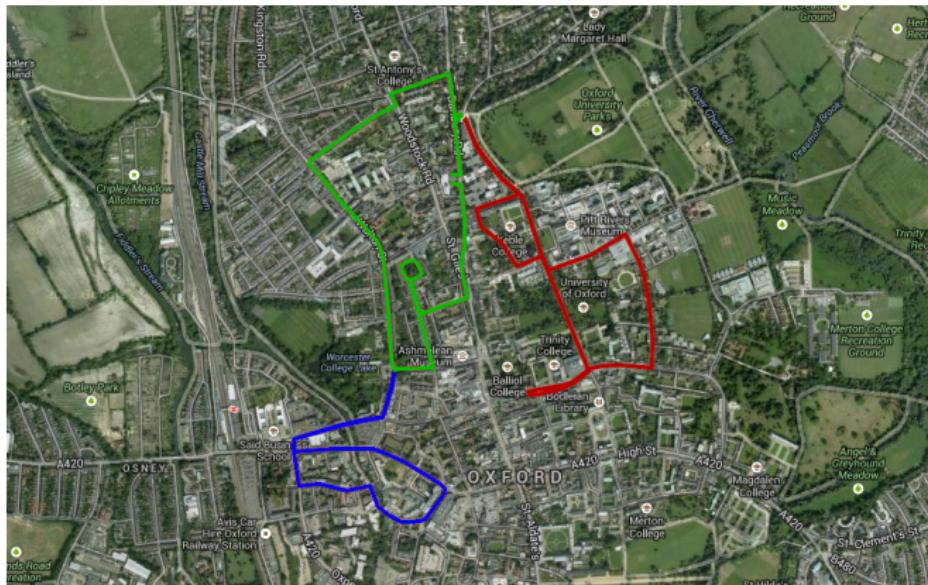
INSTITUT NATIONAL
DE L'INFORMATION
GÉOGRAPHIQUE
ET forestière

Our proposal: Learning side information with modality transfer



Experiments

Robotcar dataset



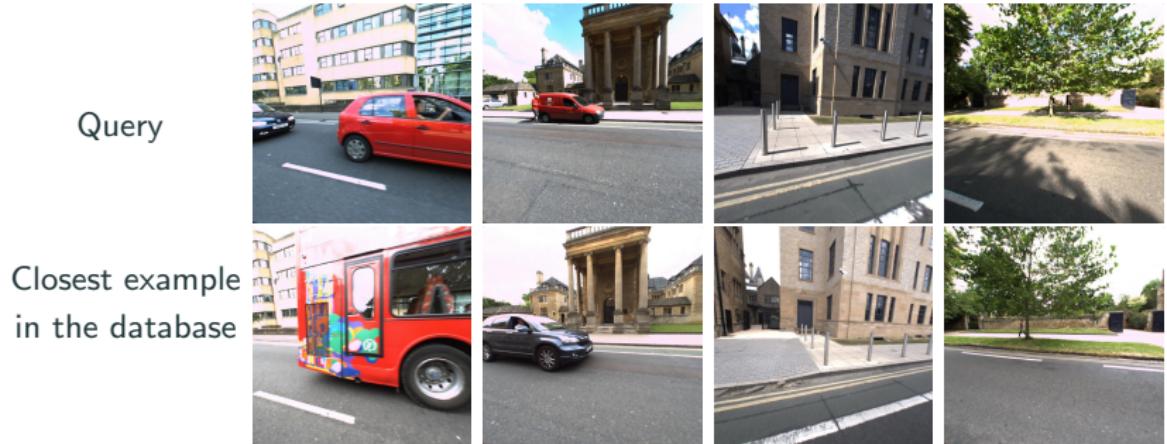
Dataset training (green), validation (blue) and testing (red) areas.



Experiments



Robotcar dataset



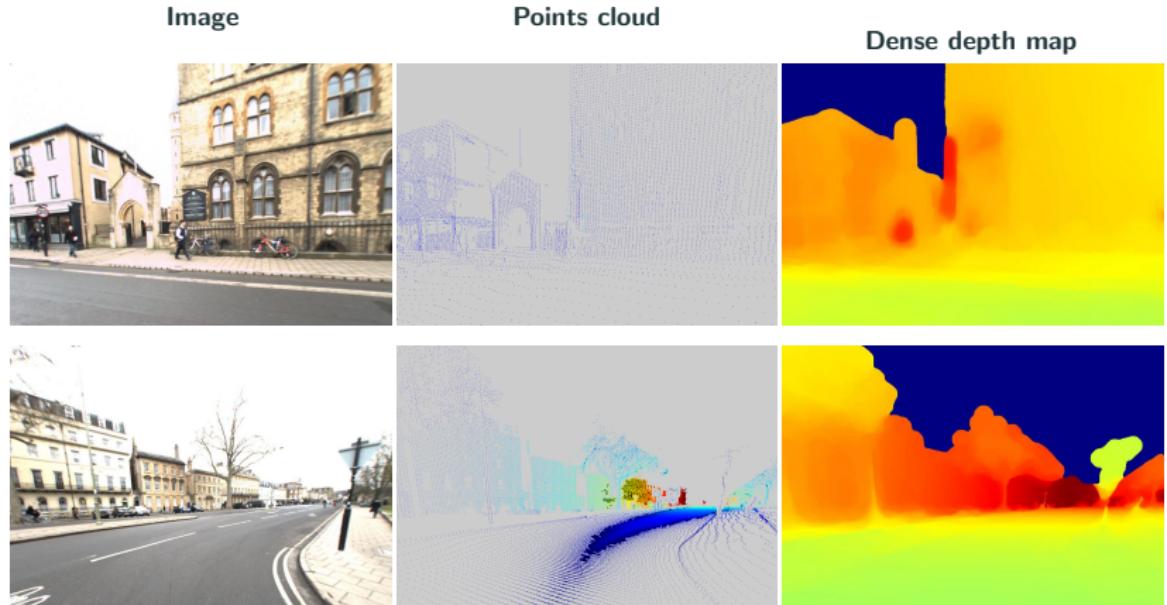
Examples of queries with corresponding dataset candidates of the testing set.

Building dense modality map



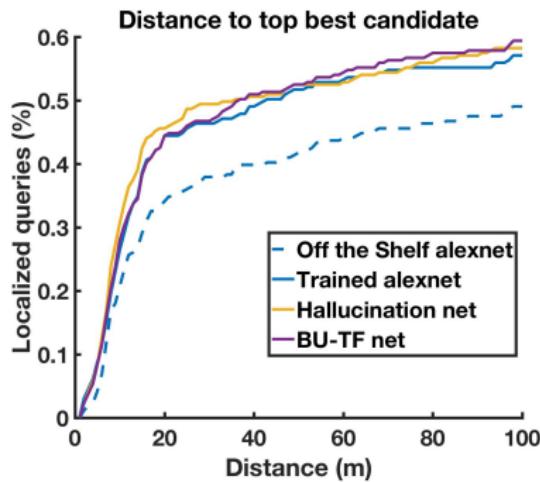
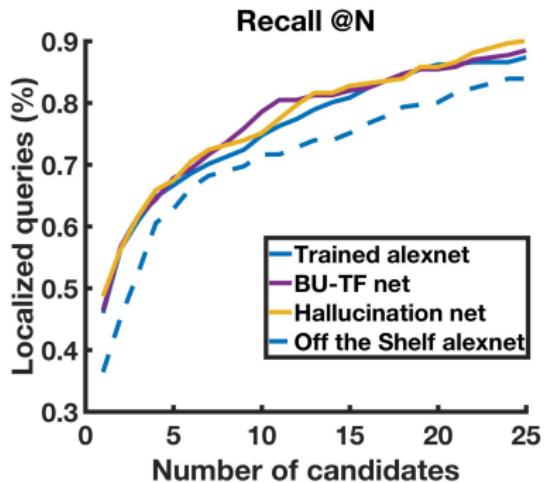
We use the algorithm proposed in [Bevilacqua et al., 2017] to create a dense modality map from an image and the associated point cloud.

Building dense modality map



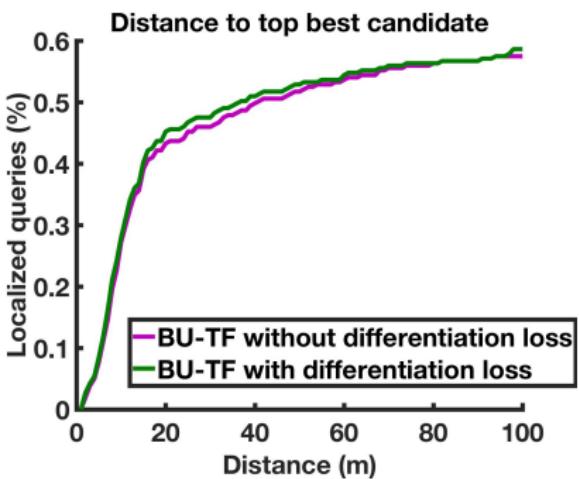
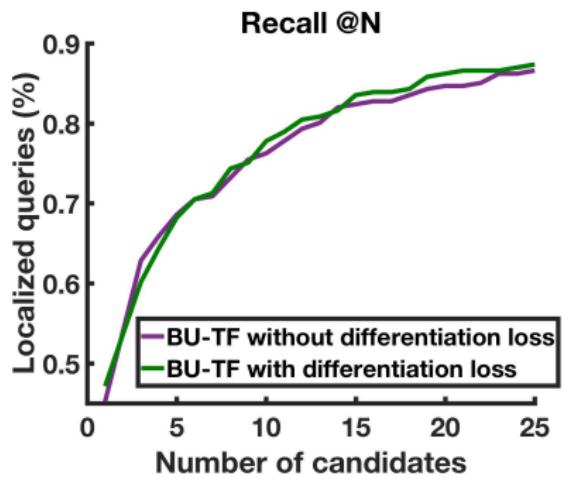
We use the algorithm proposed in [Bevilacqua et al., 2017] to create a dense modality map from an image and the associated point cloud.

Results



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

Results - Diversification loss

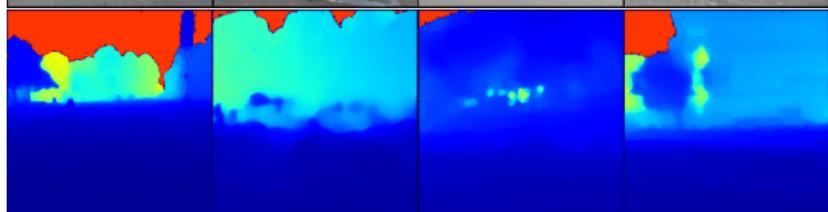


Results - Visual inspection

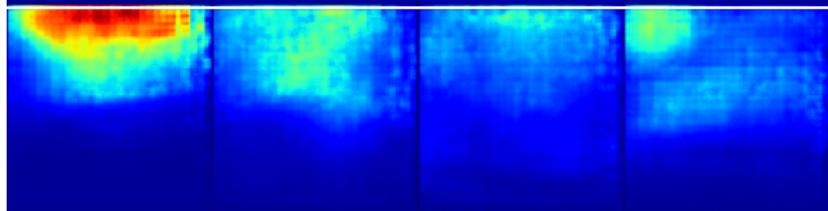
Main modality



Ground truth
side modality



Reconstructed
side modality



Experiments



Results - Conclusion

Comparison with state of the art method:

	Improvement over images-only method	Size	Training steps
Hallucination [Hoffman et al., 2016]			
BU-TF (Our proposal)			

Results - Conclusion

Comparison with state of the art method:

	Improvement over images-only method	Size	Training steps
Hallucination [Hoffman et al., 2016]	Yes		
BU-TF (Our proposal)	Yes		

Results - Conclusion

Comparison with state of the art method:

	Improvement over images-only method	Size	Training steps
Hallucination [Hoffman et al., 2016]	Yes	2x	
BU-TF (Our proposal)	Yes	1.4x	

Experiments



Results - Conclusion

Comparison with state of the art method:

	Improvement over images-only method	Size	Training steps
Hallucination [Hoffman et al., 2016]	Yes	2x	3 steps
BU-TF (Our proposal)	Yes	1.4x	2 steps

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

-  Kuznetsov, 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.
-  Radenović, F., Tolias, G., and Chum, O. (2016).
CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples.
In *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, volume 9905, pages 3–20.