

# **Visual Based Localization**

State of the Art & investigation on Modality Transfer

---

Nathan Piasco

12/10/2017

pLaTINUM project, Le Creusot

VBL State of the Art

Modality Transfer for VBL

Future research direction

Modality transfer with CNN

Preliminary Results

## **Modality Transfer for VBL**

---

**Future research direction**

## pLaTINUM context

*Find the nearest sphere in the database according to a input composed of heterogeneous data*

We want to create an indirect method to localize a query within a set of geolocalized RGB-D spheres. The considered pipeline will be:

- Create a discriminative descriptor
- Compute the representation of the spheres with our descriptor
- Compute the similarity of the query (also described by our descriptor) with the spheres representation
- Retrieve the closest sphere

## pLaTINUM context

*Find the nearest sphere in the database according to a input composed of heterogeneous data*

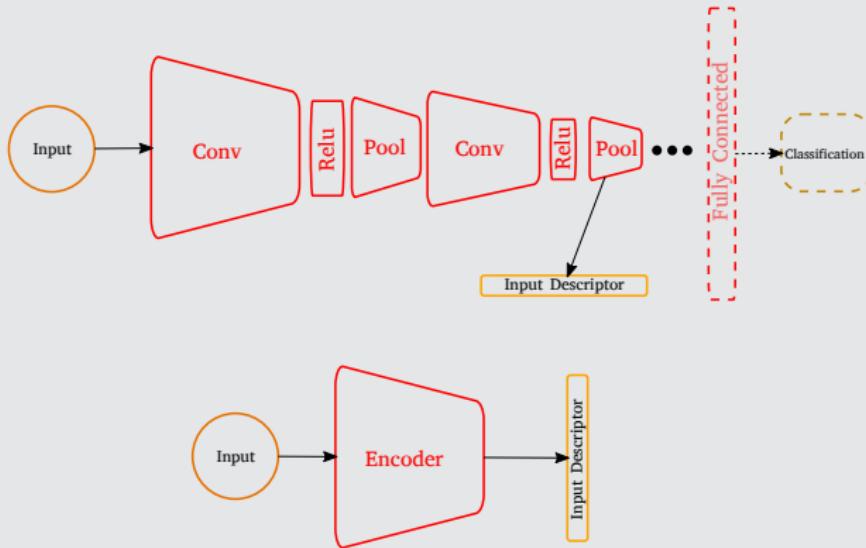
We want to create an indirect method to localize a query within a set of geolocalized RGB-D spheres. The considered pipeline will be:

- **Create a discriminative descriptor**
- Compute the representation of the spheres with our descriptor
- Compute the similarity of the query (also described by our descriptor) with the spheres representation
- Retrieve the closest sphere

# Data representation

We first focus our work on creating a robust data representation. Research on state of the art shown that CNN are the best choice.

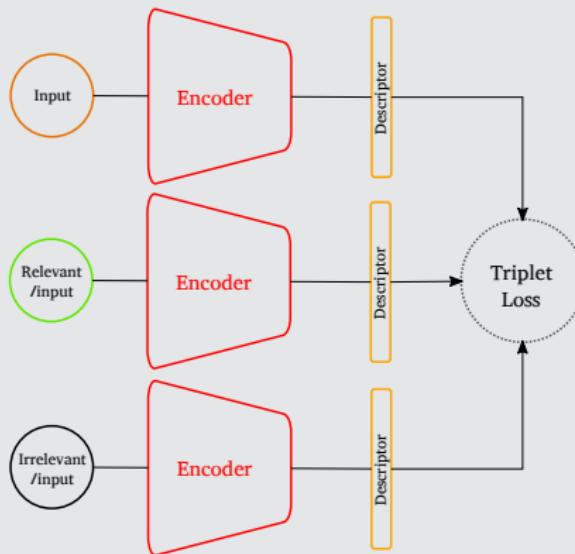
## Encoder for data representation



# Data representation

We begin with a pre-trained network and perform a fine tuning of its weights.

## Encoder training for VBL task



# Multiple modalities

How to use more than one modality at the time?

## Ideal Desired System specification

Training data type	Testing data type
RGB + Depth + Semantic + Laser reflectance + ...	RGB +/or Depth +/or Semantic +/or Laser reflectance +/or ...

## What are we first trying to do

Training data type	Testing data type
RGB + (Depth or Laser reflectance)	RGB



## **Modality Transfer for VBL**

---

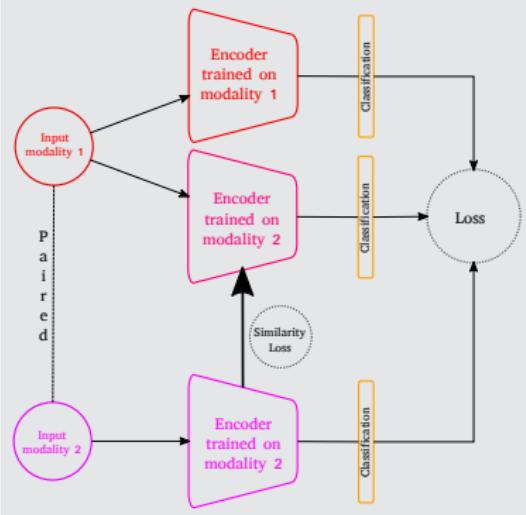
**Modality transfer with CNN**

# Modality Hallucination Network

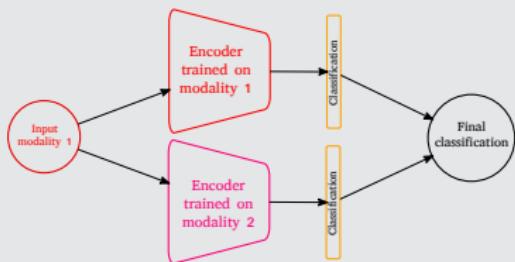
## Original contribution

From [Hoffman et al., 2016], applied for the task of semantic interpretation with Fast-RCNN.

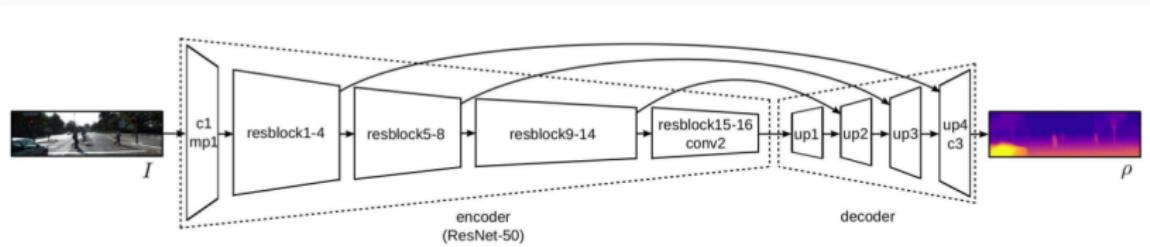
## Modality hallucination training



## The network at test time



# Depth map inference



[Kuznetsov et al., 2017]

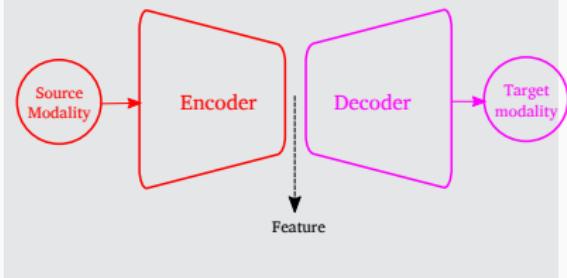
CNN can learn a model to transform a input from one modality to another (widely used to infer depth map from monocular images).

## Intuition

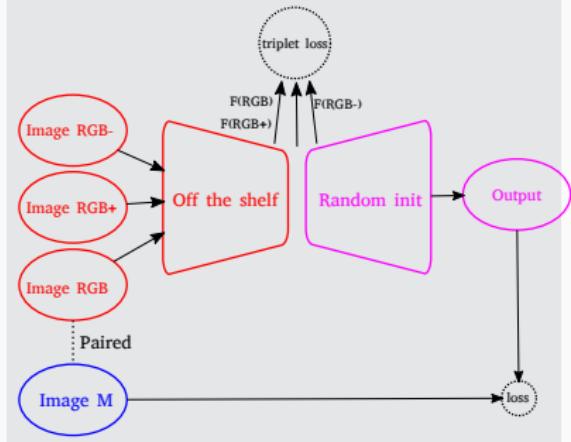
As we already use a neural network to describe our data, we should use this kind of architecture to transfer the knowledge among modalities during training.

# Proposed architecture

## Encoder-Decoder architecture



## Encoder-Decoder training scheme



Decoder part is initialized with pre-trained weights and decoder network is randomly initialized.

## **Modality Transfer for VBL**

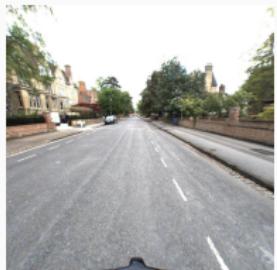
---

### **Preliminary Results**

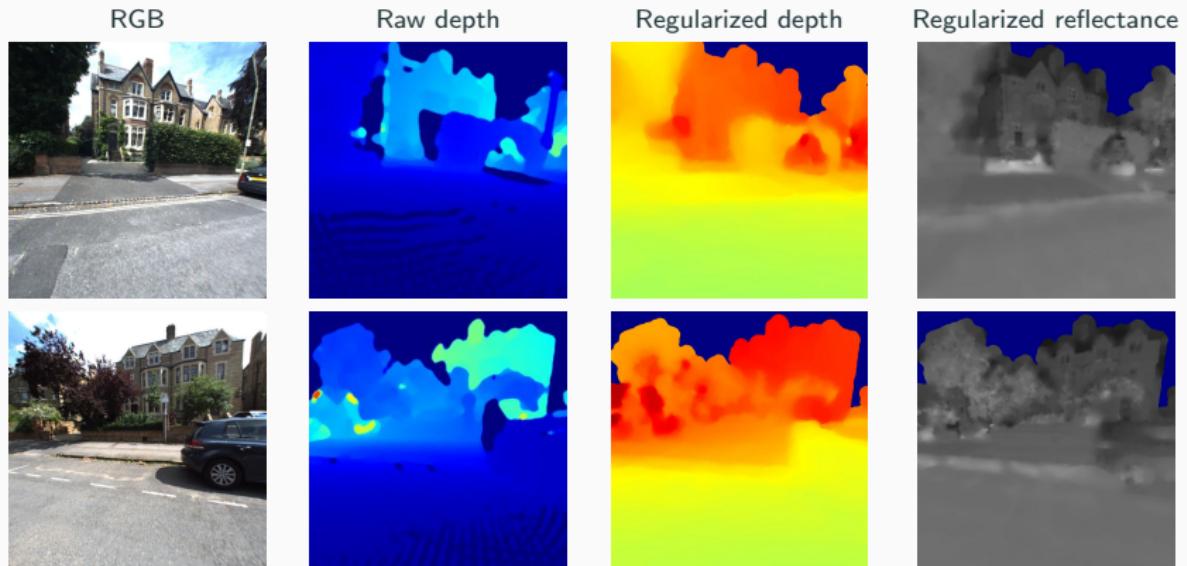
## Datasets

We use Oxford RobotCar dataset [Maddern et al., 2016] as it includes:

- Time redundancy for each car trajectory
- 4 cameras on the car & 3 LIDARS (3 modalities: RGB, Depth & Reflectance)

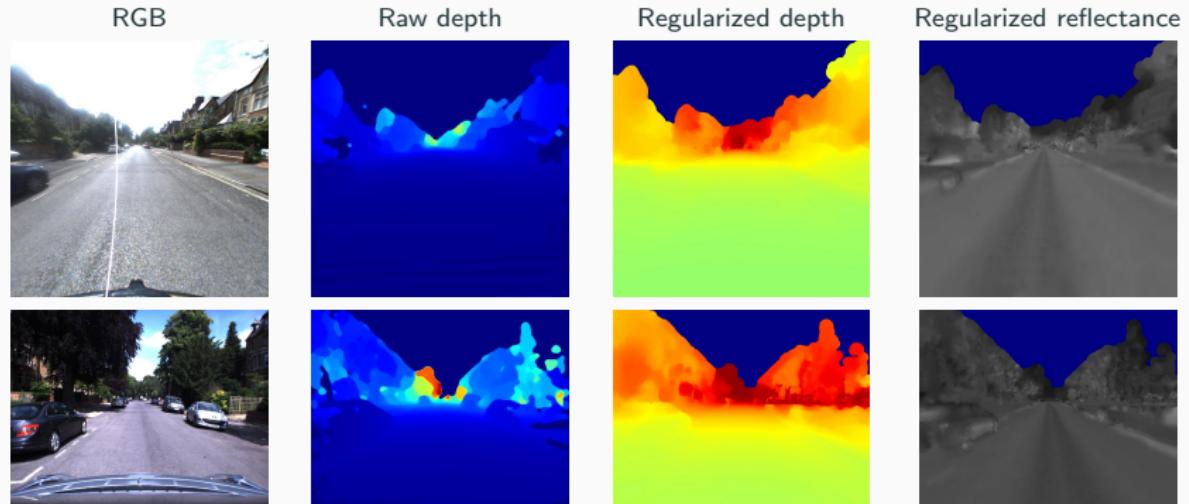


# Data



Post-processing algorithm from [Bevilacqua et al., 2017].

# Data



Data are not perfect...

## Implementation details

**Deep Learning framework:** Pytorch

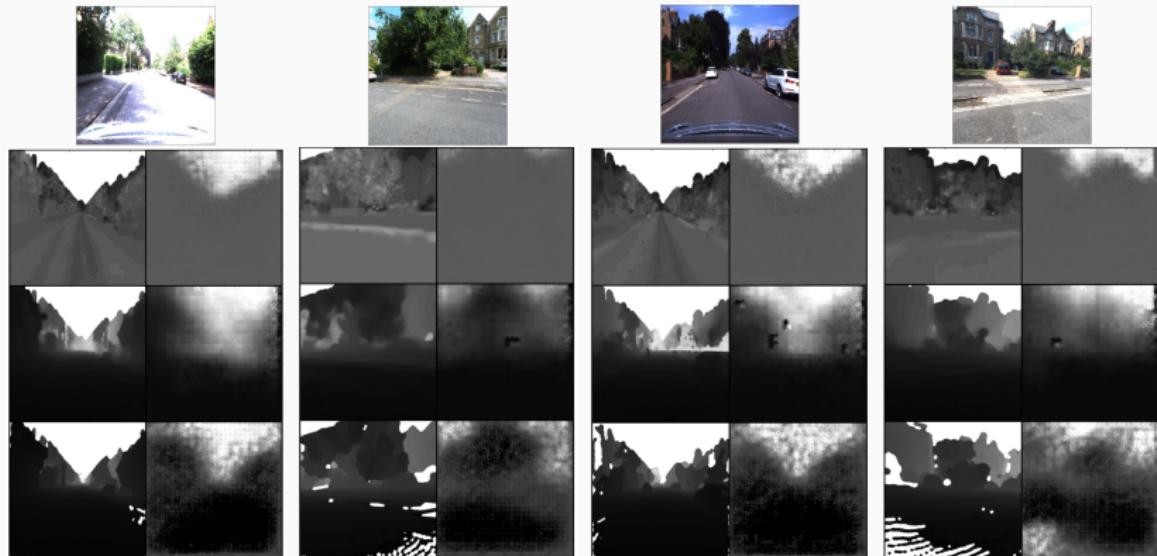
**Coder net architecture:** Alexnet

**Size of the training dataset:** 200 triplets (200 \* 3 images \* 2 modalities)

Trainings and validation frames are from a different region of Oxford than testing images.



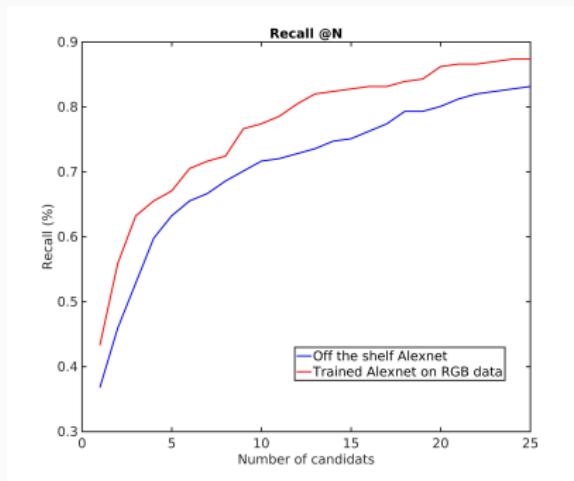
# Does the transfer learning work?



# Does it improve the localization result?

## Evaluation

We state a query as registered if one of the top  $K$  retrieved images is at a distance inferior at 25 metres.



First results **are not** better with our Encoder-Decoder net than results obtained with single Encoder training.

**Thanks for your attention**

*Discussion time*

## References I

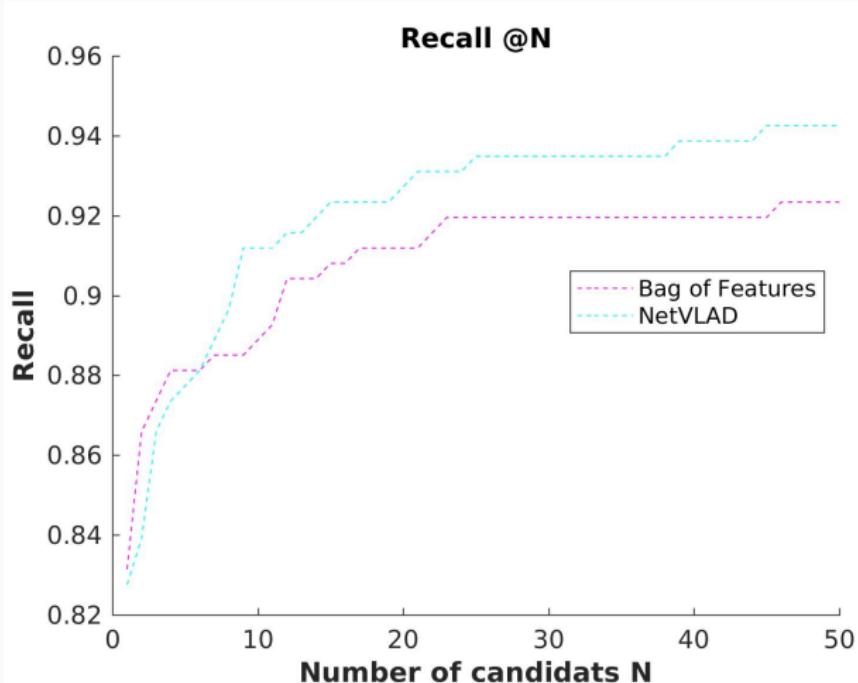
-  Arandjelović, R. and Zisserman, A. (2014).  
**DisLocation : Scalable descriptor.**  
In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.
-  Bevilacqua, M., Aujol, J. F., Biasutti, P., Brédif, M., and Bugeau, A. (2017).  
**Joint inpainting of depth and reflectance with visibility 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)*.
-  Maddern, W., Pascoe, G., Linegar, C., and Newman, P. (2016).  
**1 year, 1000 km: The Oxford RobotCar dataset.**  
*The International Journal of Robotics Research (IJRR)*, page 0278364916679498.

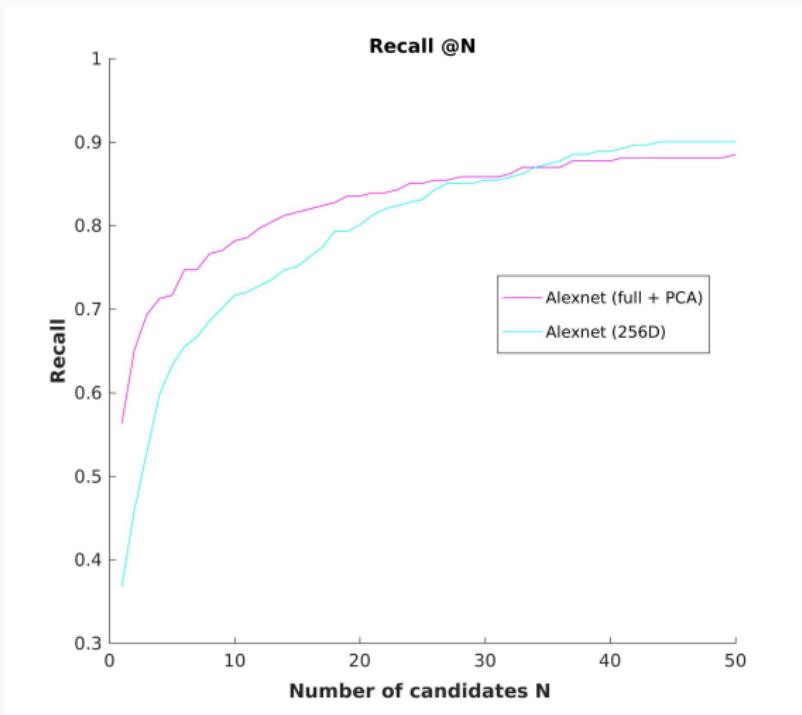
## State of the art methods

DisLocation BoF method from [Arandjelović and Zisserman, 2014].

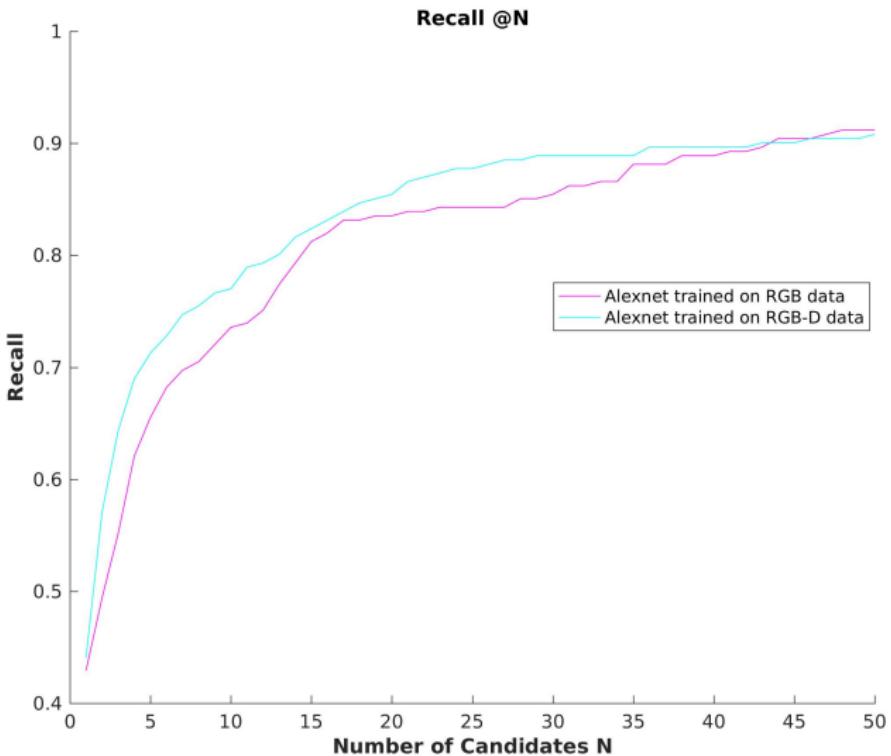


## Dimension reduction

Alexnet's descriptor is reduced in dimension with global max pooling over feature maps.



## Previous results



## Dataset trajectories

