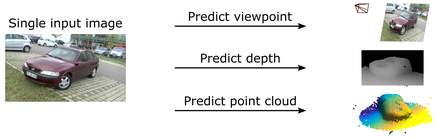
Learning 3D Object Categories by Looking Around Them

David Novotny[1,2], Diane Larlus[2], Andrea Vedaldi[1]
[1] Visual Geometry Group, University of Oxford, UK
[2] Naver Labs Europe, Grenoble, France

Introduction

Objective

> Predict the 3D structure - **Viewpoint**, **Depth** and **Point cloud** - of an unseen object instance from a **single input image**

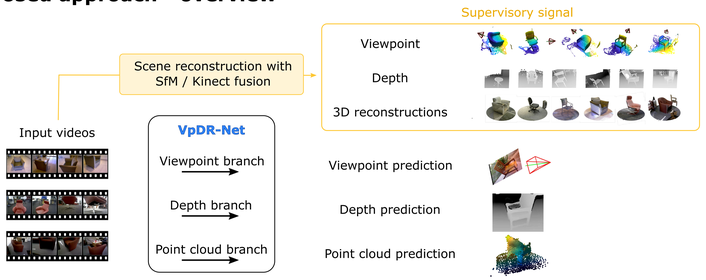


Supervision with motion

> Existing approaches [1,2] require expensive 3D annotations / synthetic dataset
> We use a weak form of supervision - **videos of object categories**

Freiburg Cars dataset (RGB videos) [3] A Large Dataset of Object Scans (RGB+D videos) [4]

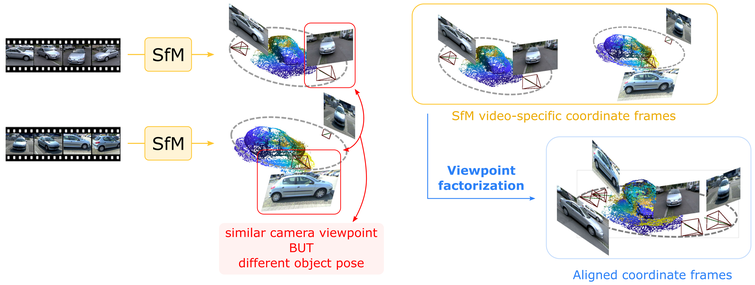
Proposed approach - overview



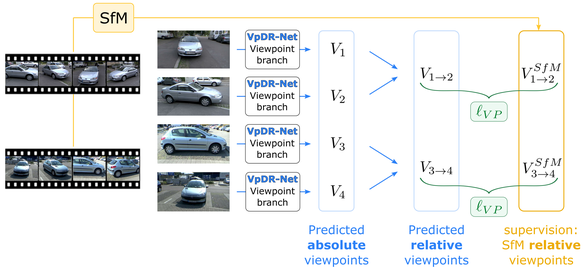
Viewpoint branch

Raw SfM absolute viewpoints cannot be used for supervising a viewpoint regressor:

Proposed solution: **Implicit viewpoint factorization**



Viewpoint factorization



Viewpoint loss ℓ_{VP}

$$\ell_{VP} = \ell_R + \ell_T$$

$$\ell_T = \|T_{12} - T^{SfM}_{12}\|^2$$

$$\ell_R = \ln \|R^{SfM}_{12} R^\top_{12}\|_F$$

$$V_{1\to2} = \{R_{12}, T_{12}\}$$

Note: We use the probabilistic extension of ℓ_{VP} (see "Confidence predictions")

Viewpoint estimation experiments

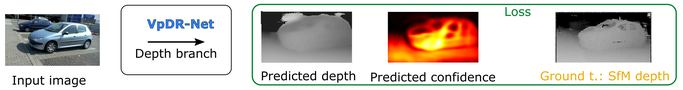
> test set: Pascal3D (cars and chairs)
> train set: Freiburg cars
  Large Dataset of Object Scans (chairs)
> benchmarked approaches:
  > VpDR-Net (viewpoint branch - implicit alignment)
  > Sedaghat et al. [3] (explicit alignment)

Camera viewpoint estimation

object class	method	↓ cam. pose error [deg]	↓ cam. center error	↓ rel. cam. pose error [deg]	↓ rel. cam. translation error
car	Sedaghat & Brox [3]	49.62	32.29	85.45	0.84
	VpDR-Net	**29.57**	**7.29**	**62.30**	**0.65**
	fully supervised	12.49	1.27	20.34	0.24
chair	Sedaghat & Brox [3]	64.68	42.46	89.01	0.95
	VpDR-Net	**42.34**	**16.72**	**71.35**	**0.93**
	fully supervised	34.37	6.14	67.41	0.74

Depth branch

> Depth prediction branch is a standard monocular depth predictor supervised with SfM / Kinect depth



Confidence predictions

Standard depth prediction loss ℓ

> l_1 distance between prediction d_{pred} and ground truth d_{gt}

$$\ell = |d_{gt} - d_{pred}|$$

> SfM depth often noisy / unreliable, we need a robust loss
> VpDR-Net needs to express uncertainty

=> solution: probabilistic loss

Probabilistic loss extension $\mathcal{L}$

> predict parameters of a laplacian distribution p_{pred} that closely follows the ground truth
> mean d_{pred} - the prediction, variance σ_{pred} - confidence

$$p_{pred} = \text{Laplace}(d_{pred} | d_{gt}, \sigma_{pred})$$

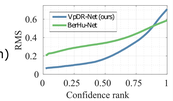
Network predictions

> the probabilistic loss $\mathcal{L}$ is the negative log-likelihood of p_{pred}

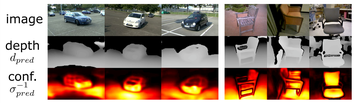
$$\mathcal{L} = -\ln p_{pred}$$

Depth estimation experiments

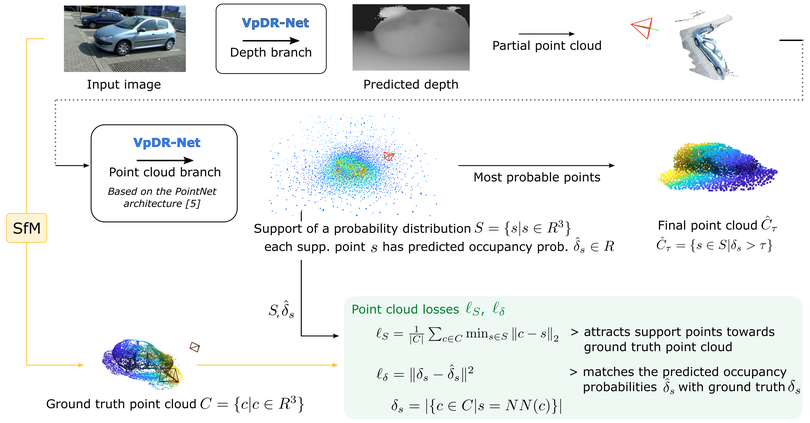
> evaluation of the confidence-based depth predictions
> test set: A Large Dataset of Object Scans
> analysis of the correlation of the confidence predictions with the actual depth error
> evaluated approaches:
  > VpDR-Net (depth branch)
  > BerHu-Net [6]



Example depth predictions



Point cloud branch



Point cloud losses ℓ_S, ℓ_δ

$$\ell_S = \frac{1}{|C|} \sum_{c \in C} \min_{s \in S} \|c - s\|_2$$

> attracts support points towards ground truth point cloud

$$\ell_\delta = \|\delta_s - \hat{\delta}_s\|^2$$

$$\delta_s = |\{c \in C | s = NN(c)\}|$$

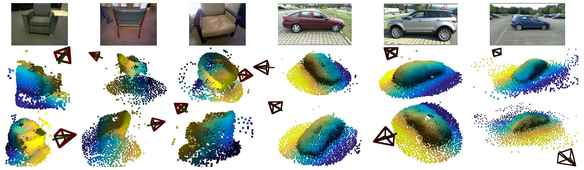
> matches the predicted occupancy probabilities $\hat{\delta}_s$ with ground truth δ_s

Point cloud prediction experiments

> test set: Freiburg cars / A Large Dataset of Object Scans (chairs)
> train set: Freiburg cars / A Large Dataset of Object Scans (chairs)
> benchmarked approaches:
  > VpDR-Net (point cloud branch)
  > Aubry et al. [7] (3D model retrieval approach)

Quality of the predicted point clouds

object class	chair		car	
metric	↑ voxel IoU	↓ pcl distance	↑ voxel IoU	↓ pcl distance
Aubry et al. [7]	0.06	1.30	0.21	0.41
VpDR-Net	**0.13**	**0.19**	**0.26**	**0.26**



Conclusions

> Proposed **VpDR-Net** - a video-supervised CNN for monocular prediction of viewpoint, depth and point cloud of an object
> VpDR-Net outperforms baselines on all three tasks
> Main contributions:
  > **Viewpoint factorization** network for learning absolute viewpoint from misaligned SfM reconstructions
  > **Probabilistic losses** for robust optimization and expression of prediction confidence in deep networks
  > **Point cloud completion** network for dealing with noisy and sparse SfM point cloud supervision

References

[1] Tulsiani and Malik: Viewpoints and Keypoints, CVPR 2015
[2] Su et al.: Render for CNN: Viewpoint Estimation in Images Using CNNs Trained with Rendered 3D Model Views, ICCV 2015
[3] Sedaghat and Brox.: Unsupervised Generation of a Viewpoint Annotated Car Dataset from Videos, ICCV 2015
[4] Choi et al.: A Large Dataset of Object Scans, arXiv
[5] Qi et al.: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation, CVPR 2017
[6] Laina et al.: Deeper Depth Prediction with Fully Convolutional Residual Networks, 3DV 2016
[7] Aubry et al.: Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models, CVPR 2014