DeFeat-Net: General Monocular Depth via Simultaneous Unsupervised Representation Learning

CVSSP Centre for Vision, Speech and Signal Processing

Jaime Spencer Martin (jaime.spencer@surrey.ac.uk) **Richard Bowden (r.bowden@surrey.ac.uk)** **Simon Hadfield (s.hadfield.@surrey.ac.uk)**

Centre for Vision, Speech and Signal Processing, Faculty of Engineering and Physical Sciences, University Of Surrey, Guildford, Surrey, GU2 7XH

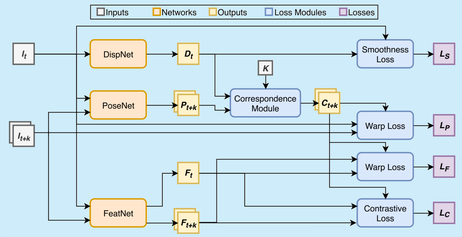
UNIVERSITY OF SURREY

1 - Abstract

- We present **DeFeat-Net**, a multi-task approach to robust depth + feature learning
- Current **photometric-based losses break down** in dimly-lit environments
- Incorporating an **additional feature learning** task improves nighttime robustness, whilst still allowing for **fully unsupervised** training
- **Code** available at **github.com/jspenmar/DeFeat-Net**

Eval Sample *Monodepth v2* *Proposed*

2 - Overview



3 - Networks

- All networks use a separate ResNet18 encoder
- DepthNet
  - Convolutional decoder with **skip connections**
  - Produces a **dense disparity** map
  - **Normalized** between [0, 1] and rescaled to desired depth range
- FeatNet
  - Convolutional decoder with **skip connections**
  - Produces a **dense n-dimensional** feature map, i.e. (H x W x n)
  - Features are **L2 normalized**
- PoseNet
  - **6DoF** pose regression
  - **Normalized** translation, rotation as **axis-angle**

4 - Correspondence Module

- From **depth + motion**, a set of **correspondences** between images can be obtained
- Correspondences are filtered using the **minimum reprojection** error and **automasking**

$$q = \pi^{-1}(\dot{p}) = K_t^{-1}\dot{p}\, D_t(p)$$

$$c_{t\to t+k}(p) = \pi(\dot{q}) = K_t P_{t\to t+k}\dot{q}$$

- Uses:
  - **Bilinear sampling** from images/features in photometric loss

5 - Losses

- Pixel-wise Contrastive
  - Contrastive loss to drive **feature learning** and separation
  - **Positive samples** are given by the obtained **correspondences**
  - **Negative samples** are obtained by **randomly sampling** image locations

$$l(y, f_1, f_2) = \begin{cases} \frac{1}{2}(d)^2 & \text{if } y = 1 \\ \frac{1}{2}\{\max(0, m-d)\}^2 & \text{if } y = 0 \\ 0 & otherwise \end{cases}$$

- Photometric Warp
  - Base photometric loss
  - Weighted combination of **SSIM and L1** losses
  - Evaluates the **image reconstruction** from correspondences

- Feature Warp
  - Similar to Photometric Warp, but applied to the **dense features**
  - Weighted combination of **SSIM and L1**
  - Evaluates **dense feature reconstruction** from correspondences

- Smoothness Loss
  - Discourages changes in depth unless there's an edge in the image
  - **Prevents smoothing** over edges

$$L_S = \frac{\lambda}{N}\sum_p |\partial D_t(p)|\, e^{-||\partial I_t(p)||}$$

7 - Evaluation

7.1 - Single Domain

- Training & evaluation on single domain daytime data: **Kitti**
- Competitive results with current SOTA

- Depth: Evaluation on **Eigen-Zhou split**

Method	Abs-Rel	Sq-Rel	RMSE	RMSE-log	A1	A2	A3
LEGO [75]	0.162	1.352	6.276	0.252	-	-	-
Ranjan [54]	0.148	1.149	5.464	0.226	0.815	0.935	0.973
EPC++ [42]	0.141	1.029	5.350	0.216	0.816	0.941	0.976
Struct2depth (M) [8]	0.141	1.026	5.291	0.215	0.816	0.945	0.979
Monodepth V2 [22]	**0.123**	0.944	5.061	**0.197**	**0.866**	**0.957**	**0.980**
DeFeat	0.126	**0.925**	**5.035**	0.200	0.862	0.954	**0.980**

Kitti depth evaluation

- Features: **Classification AUC** and **average distance** for positives and negatives
- Sample negatives from the whole image (**Global**) or 25-pixel radius (**Local**)

Method	μ_+	Global μ_-	Global AUC	Local μ_-	Local AUC
ORB [56]	N/A	N/A	85.83	N/A	84.06
ResNet [26]	8.5117	25.9872	94.77	11.1335	68.26
ResNet-L2	0.341	1.0391	99.25	0.4371	71.80
VGG [64]	4.0077	12.6543	92.94	5.9088	70.03
VGG-L2	0.3905	1.2235	99.57	0.565	77.06
SAND-G [65]	**0.093**	0.746	**99.73**	0.266	87.06
SAND-L	0.156	0.592	98.88	0.505	**94.34**
SAND-GL	0.183	0.996	99.28	0.642	93.34
DeFeat	0.105	1.113	99.10	0.294	83.64

Kitti feature evaluation

7.2 - Multi-Domain

- Trained on **RobotCar Seasons**
  - Data from multiple seasons: **day**, **night**, snow, rain, overcast…
  - **No ground truth** depth/correspondence data

- Depth
  - Evaluate on **original RobotCar** subset (6k day, 6k night)
  - **Outperforms current SOTA** (also trained on RobotCar Seasons)
  - For nighttime data, photometric consistency assumptions break down
  - **Feature learning + warping** provides robust/**strong supervision**

Test domain	Method	Abs-Rel	Sq-Rel	RMSE	RMSE-log	A1	A2	A3
Day	Monodepth V2 [22]	0.271	3.438	9.268	0.329	**0.600**	0.840	0.932
Day	**DeFeat**	**0.265**	**3.129**	**8.954**	**0.323**	0.597	**0.843**	**0.935**
Night	Monodepth V2 [22]	0.367	4.512	9.270	0.412	0.561	0.790	0.888
Night	**DeFeat**	**0.335**	**4.339**	**9.111**	**0.389**	**0.603**	**0.828**	**0.914**

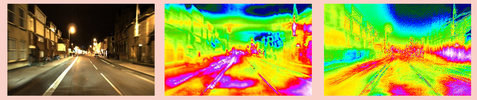
RobotCar depth evaluation

7 - Evaluation cont'd

7.3 - Multi-Domain cont'd

Eval Sample *Monodepth v2* *DeFeat-Net*

- Features
  - **No ground truth** correspondences available
  - **Visualize** features with **linear projection** showing image correlation



Eval Sample *Single Domain* *Multi-Domain*

7.3 - Ablation

- Explore the benefit of **concurrent training**
- Re-train on each dataset, **disabling** the **FeatNet** subsystem
- Results show how, especially for night-time data, **feature learning is a crucial component**

Dataset	Method	Abs-Rel	Sq-Rel	RMSE	RMSE-log	A1	A2	A3
KITTI	DeFeat (no feat)	**0.123**	0.948	5.130	**0.197**	**0.863**	**0.956**	**0.980**
KITTI	**DeFeat**	0.126	**0.925**	**5.035**	0.200	0.862	0.954	**0.980**
RobotCar Day	DeFeat (no feat)	0.274	3.885	**8.953**	0.335	**0.640**	**0.853**	0.934
RobotCar Day	**DeFeat**	**0.265**	**3.129**	8.954	**0.323**	0.597	0.843	**0.935**
RobotCar Night	DeFeat (no feat)	0.748	13.502	**8.956**	0.657	0.393	0.624	0.759
RobotCar Night	**DeFeat**	**0.335**	**4.339**	9.111	**0.389**	**0.603**	**0.828**	**0.914**

Ablation results per dataset

8 - Conclusions

- We introduce **DeFeat-Net**, a multi-task learning approach to unsupervised **depth**, **motion** and **dense feature** learning
- The incorporation of a novel feature learning task improves **depth estimation in adverse conditions**
- This is achieved by providing an **additional feature warp** loss, which is robust to appearance changes
- Future work could attempt to enforce **consistency** across **multiple seasons**

This work was partially funded by the EPSRC under grant agreements (EP/R512217/1, EP/S016317/1 and EP/S035761/1). We would also like to thank NVIDIA Corporation for their Titan Xp GPU grant.

EPSRC Engineering and Physical Sciences Research Council