Test-Time Adaptation for Depth Completion

Hyoungseob Park
hyoungseob.park@yale.edu
Yale Vision Lab

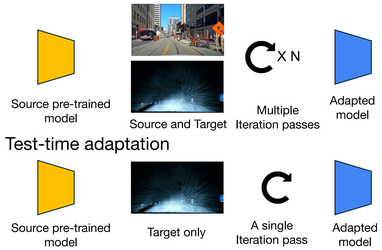
Anjali Gupta
anjali.gupta@yale.edu
Yale Vision Lab

Alex Wong
alex.wong@yale.edu
Yale Vision Lab

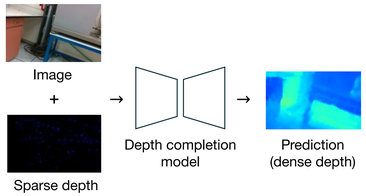
Paper | Code | Project page

Preliminaries

Unsupervised domain adaptation



Depth completion

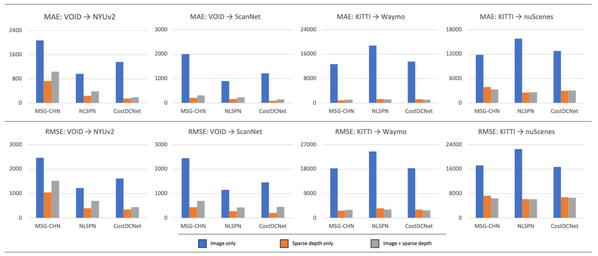


Test-time adaptation: adaption of a source pre-trained model to the target domain within a single epoch, **without accessing the source data**

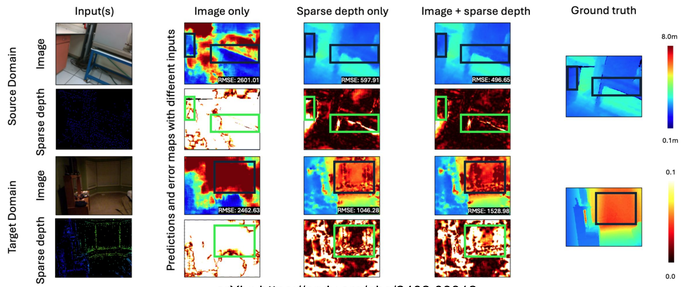
Depth completion: estimate **a dense depth map** from image and its synchronized sparse range (e.g., LiDAR, SLAM)

Motivation

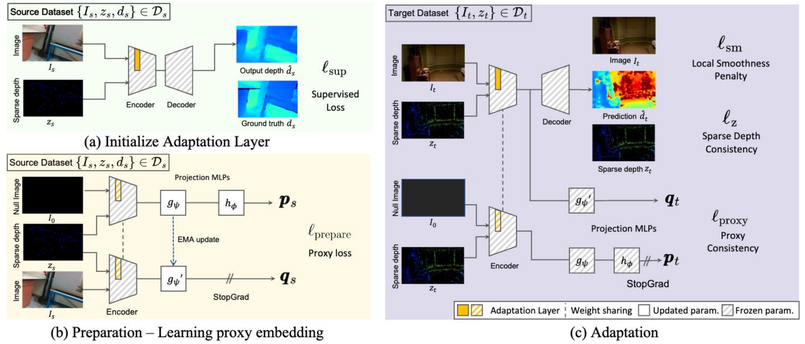
Quantitative and qualitative results of pretrained models when one of its input modality is absent. Sparse range is less sensitive to domain shift in multimodal depth estimation models.



Forgoing the image modality at test-time yields similar or better results as using both modalities.



Methodology



ProxyTTA deploys a single adaptation layer and leverages sensor modalities that are ***less sensitive*** to test-time nuisances to predict "proxy" (of the source domain) embeddings for guiding the model update to map test-time domain features back to source domain for other *more sensitive* modalities

Before deployment

In stage (a), Adaptation Layer is implemented on RGB branch and initialized by supervised loss.

In stage (b), Projection MLPs learns proxy embedding by minimizing the distance between sparse-range embedding and RGB image + sparse range embedding.

$$\ell_{\text{prepare}} = 1 - \left(\frac{\boldsymbol{p}_s}{\|\boldsymbol{p}_s\|} \cdot \frac{\boldsymbol{q}_s}{\|\boldsymbol{q}_s\|}\right) \qquad \begin{aligned} \boldsymbol{p}_s &= h_\omega(g_\psi(\text{StopGrad}(e_\phi(I_0, z_s)))), \\ \boldsymbol{q}_s &= \text{StopGrad}(g_{\psi'}(e_\phi(I_s, z_s))) \end{aligned}$$

After deployment

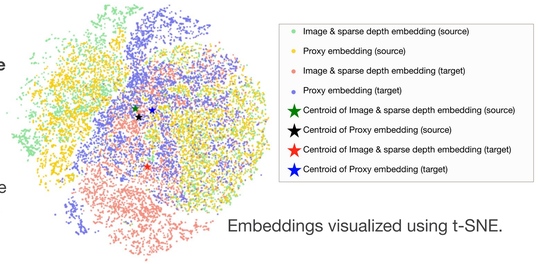
In stage (c), Adaptation layer is updated by sparse depth loss, smoothness loss, and proxy loss

$$\mathcal{L}_{\text{adapt}} = \underbrace{w_z \ell_z + w_{\text{sm}} \ell_{\text{sm}}}_{\text{Data term}} + \underbrace{w_{\text{proxy}} \ell_{\text{proxy}}}_{\text{Proxy term}} \qquad \ell_{\text{proxy}} = 1 - \left(\frac{\boldsymbol{p}_t}{\|\boldsymbol{p}_t\|} \cdot \frac{\boldsymbol{q}_t}{\|\boldsymbol{q}_t\|}\right) \qquad \begin{aligned} \boldsymbol{p}_t &= \text{StopGrad}(h^*_\omega(g^*_\psi(e_\phi(I_0, z_t)))) \\ \boldsymbol{q}_t &= g^*_{\psi'}(e_\phi(I_t, z_t)). \end{aligned}$$

Analysis

Embeddings encoding **image and sparse range from the source (train) domain** are far from **image and sparse range from the source domain from the target (test) domain**.

Proxy embeddings from just sparse range modality in target domain are close to embeddings encoding both the image and sparse from source domain.

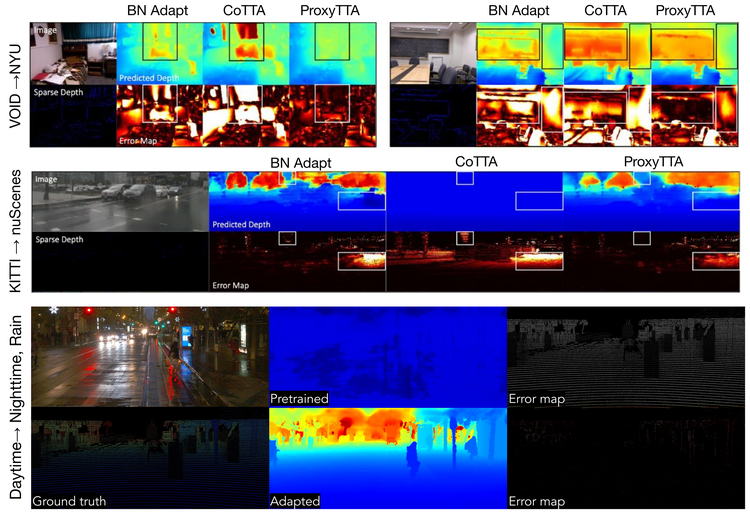


Embeddings visualized using t-SNE.

Results

ProxyTTA improved Baseline by 44%, CoTTA by 26%, BN Adapt by 16%

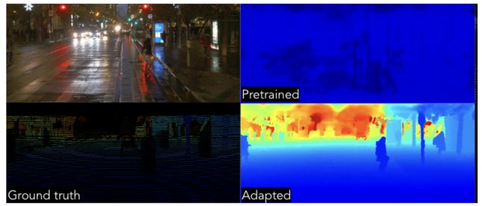
Method		MAE	RMSE	MAE	RMSE	MAE	RMSE
		KITTI → VKITTI-FOG		KITTI → nuScenes		KITTI → Waymo	
MSG-CHN	Pretrained	2842.88	6557.38	3331.821	6449.094	1107.22	2962.45
	CoTTA	*730.6±11.67*	*3330.23±44.83*	*3157.69*	*6434.14*	*655.77±30.98*	*2213.27±98.80*
	ProxyTTA-fast (Ours)	**728.24±3.73**	**3087.36±15.92**	**2834.08±17.64**	**6096.56±21.08**	**608.91±1.74**	**1921.83±2.54**
NLSPN	Pretrained	1309.99	7423.48	2656.609	6146.590	1175.83	3078.377
	BN Adapt, ℓ_z, ℓ_{sm}	775.20±5.65	3465.05±32.73	2928.51±75.89	8209.24±164.31	*494.94±3.08*	*1921.17±338.06*
	CoTTA	767.93±5.47	3799.88±17.29	*2650.45±15.04*	6242.52±33.14	933.41±4.31	2763.88±143.48
	ProxyTTA (Ours)	**686.91±22.14**	**2666.70±56.64**	**2589.25±59.03**	**6006.18±90.66**	**477.28±3.32**	**1598.64±18.95**
CostDCNet	Pretrained	1042.98	6301.60	3064.724	6630.649	1093.79	2798.25
	BN Adapt, ℓ_z, ℓ_{sm}	*729.67±3.14*	3413.76±14.59	*2288.85±14.02*	6338.38±31.31	*469.97±2.47*	**1572.95±10.63**
	CoTTA	756.32±3.59	3686.69±14.75	2676.83±68.92	*6099.49±66.79*	689.94±1.95	2140.23±16.12
	ProxyTTA (Ours)	**512.72±0.74**	**2735.01±3.53**	**2062.28±11.24**	**5509.96±23.41**	**466.44±1.63**	*1580.38±11.48*
		VOID → NYUv2		VOID → SceneNet		VOID → ScanNet	
MSG-CHN	Pretrained	1040.934	1528.983	281.28	645.01	687.988	1201.747
	CoTTA	*876.93±146.95*	*1148.62±173.53*	*223.19±14.77*	*498.46±28.21*	*619.37±4.14*	*1141.04±7.35*
	ProxyTTA-fast (Ours)	**699.60±6.00**	**1120.37±9.76**	**192.74±1.72**	**424.49±4.58**	**302.21±4.10**	**480.08±8.03**
NLSPN	Pretrained	388.87	702.80	167.250	438.71	233.33	431.20
	BN Adapt, ℓ_z, ℓ_{sm}	*147.55±1.36*	*271.10±2.17*	*120.48±1.94*	*345.91±7.14*	*82.76±0.47*	*181.97±1.21*
	CoTTA	390.50±8.29	704.72±16.74	205.02±1.79	540.01±4.08	234.77±1.52	496.18±2.75
	ProxyTTA (Ours)	**124.41±2.27**	**240.73±5.72**	**113.93±1.49**	**333.41±4.32**	**74.77±0.31**	**166.61±0.45**
CostDCNet	Pretrained	189.10	446.71	173.37	443.22	144.31	458.69
	BN Adapt, ℓ_z, ℓ_{sm}	136.80±5.35	338.59±22.36	385.9±6.68	*68.44±0.46*	*164.59±2.82*
	CoTTA	147.69±5.3	376.87±21.25	136.42±3.41	405.38±11.63	101.98±1.53	322.63±5.04
	ProxyTTA (Ours)	**95.87±2.16**	**203.83±4.72**	**125.75±1.93**	*357.12±4.13*	**68.17±0.44**	**162.35±1.12**



arXiv: https://arxiv.org/abs/2402.03312

github: https://github.com/seobbro/TTA-depth-completion

This work was supported by NSF 2112562 (Athena AI Institute).


Pretrained
Ground truth
Adapted

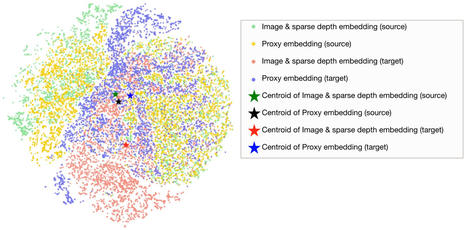


Image & sparse depth embedding (source)
Proxy embedding (source)
Image & sparse depth embedding (target)
Proxy embedding (target)
Centroid of Image & sparse depth embedding (source)
Centroid of Proxy embedding (source)
Centroid of Image & sparse depth embedding (target)
Centroid of Proxy embedding (target)