Turegali Amir
master degree student al-Farabi Kazakh National University,
Kazakhstan, c. Almaty

CROSS-ATTENTIVE PROXY REFINEMENT FOR DEPTH COMPLETION

Models for depth completion often suffer from significant performance degradation when deployed in environments different from those they were trained on, due to domain shifts in image appearance and scene structure. While prior work has attempted to address this challenge using test-time adaptation, existing methods either rely on frozen source-trained embeddings or adapt only a limited subset of model parameters. In this paper, we propose CAPR (Cross-Attentive Proxy Refinement for Depth Adaptation), a novel test-time adaptation framework that dynamically aligns RGB features to the source domain using sparse depth as a stable geometric prior. CAPR introduces two vations: (1) a Cross-Attentive Proxy Fusion (CAPF) module that generates instance-aware proxy embeddings by conditioning RGB features on sparse depth, (2) Self-Regularizing Dual Adaptation (SDA) layers that co-adapt both RGB and sparse depth branches to achieve modality-consistent alignment, CAPR requires no access to source data or labels during test-time and adapts efficiently in a single pass. Extensive experiments on standard indoor and outdoor benchmarks demonstrate that CAPR outperforms strong baselines including ProxyTTA, CoTTA, and BN Adapt, achieving up to 21.3% RMSE reduction while preserving low runtime overhead. Our results establish CAPR as a robust and scalable solution for depth completion under domain shift

Key Words: Depth completion, Domain Adaptation, Test-Time Adaptation, Multimodal Fusion.

Related work: Convolutional Spatial Propagation Network (CSPN [1]) and NLSPN [2]. Li et al. proposed a CSPN method in their paper, in which the initial depth map is refined using trainable convolution kernels and the information is propagated over a local window. NLSPN extends this idea to a non-local domain, for which an affinity matrix is trained to connect non-local points, thereby allowing depth adjustments over large distances, providing high edge accuracy and improving detail

of small objects. However, such gradual refinement requires additional memory to store affinity maps and increases the running time of the algorithm. Multi-scale Cascaded Hourglass [3] Park et al. proposed a cascade of hourglass modules, where each subsequent module receives the output of the previous one as input, thereby refining the depth map at a different scale. Each hourglass consists of downward and upward convolutions with skip connections. For each stage, such a cascade system allows the network to sequentially correct large-scale approximation errors and restore details on small objects, but the requirements for tuning the learning rate and loss coefficients are higher. Cost Volume Based Networks [4] (CostDCNet). CostDCNet type networks build cost-volume by disparity, stereo calculation methods work on a similar principle by performing 3D convolution by volume to select the most probable depth value. This method, like the previous one, requires significant computing resources due to the processing of three-dimensional volume, but the approach works well in areas with ambiguous depth solutions, for example, on flat surfaces.

Test-Time Adaptation for Depth Completion The deployment of depth completion models trained on source datasets to target domains results in performance degradation which Test-Time Adaptation [5] (TTA) addresses. The goal of TTA methods is to modify models during testing through test data analysis without needing source data access or multiple test-time iterations. The conventional TTA approaches modify particular model components including batch normalization layers and classification heads through objectives that use entropy minimization or self-supervised tasks. The developed approaches function best for classification and segmentation tasks but show reduced effectiveness in regression tasks including depth completion. ProxyTTA [6] represents a new TTA technique which was created to work with depth completion tasks. The sparse depth input shows reduced sensitivity to domain shift variations compared to RGB images. During source domain training ProxyTTA develops MLP-based mappings that transform sparse depth features into joint image-depth features to create proxy embeddings which represent source-domain multimodal features. During testing the proxy embeddings from ProxyTTA serve as guidance for adapting an auxiliary adaptation layer which gets inserted into the image

encoder branch. The method enhances performance by maximizing the similarity between test-time image-depth features and proxy embeddings to match target domain image modality with source domain features. The method includes additional regularization through sparse depth consistency and local smoothness losses while it updates only the adaptation layer to maintain efficiency.

Proposed method: CAPR (Cross-Attentive Proxy Refinement for Depth Adaptation) propose to replace the static MLP projection of proxy embeddings in ProxyTTA with a dynamic multi-head cross-attention mechanism which allows for more flexible "gluing" of RGB and deep features at the test adaptation stage. The general scheme of CAPR operation is as follows:

1. At the input we have two streams of features:

$$F_{RGB} \in R^{H*W*C} \tag{1}$$

$$F_{Depth} \in R^{H*W*C} \tag{2}$$

2. Transform them into sequences (flattening) to apply multi-head cross-attention:

$$Attention(Q,\ K,\ V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where

$$Q(query) = F_{RGB} * W_Q \tag{4}$$

$$K(key), V(value) = F_{Depth} * W_K, W_V \tag{5}$$

W_Q, W_K, W_V — training matrices, d_k — the dimension of one attention-head.

3. LoRA [7] is used inside attention: training matrices W_Q, W_K, W_V are modified as

$$W = W_0 + \alpha A * B, \tag{6}$$

$$A \in R^{d*r}, B \in R^{r*d}$$

$r \ll d$.low rank gives economy of parameters; W_0 — frozen weights from pretrained model; α — scale for LoRA component.

4. Real F_{RGB} go through the adaptation layer and train it so that the output of the adaptation layer is as close as possible to $\hat{F}_{RGB}$ in cosine proximity:

$$L_{proxy} = 1 - cos\,(Adapted\,RGB, \hat{F}_{RGB}) \quad (7)$$

5. Instead of a static template, a dynamically generated, scene-dependent reference is specified. To avoid unnecessary adaptation, we introduce a moving average strategy with two thresholds. The logic of two-way hysteresis with two thresholds τ_{high} and τ_{low} ($\tau_{high} > \tau_{low}$) is used to decide whether to skip or perform adapt steps on the next frame. When $\bar{l}_t$ exceeds τ_{high}, this indicates a significant shift in the domain, and for frame *t+1*, adaptation is enabled. If $\bar{l}_t$ falls below τ_{low}, then high accuracy of the model is maintained without additional adaptation, and on the next frame gradient-step is skipped. For $\bar{l}_t$ values between these thresholds, the "adapt/not adapt" state does not change and preserves the previous one, which prevents multiple switches due to minor loss fluctuations.

$$adapt = \begin{cases} false,\ \bar{l}_t > \tau_{high} \\ true, \bar{l}_t < \tau_{low} \\ adapt, \tau_{low} < \bar{l}_t < \tau_{high} \end{cases}$$

Results: In indoor scenarios (VOID → NYUv2, SceneNet, ScanNet), CAPR also demonstrates competitive results. For example, on the NYUv2 dataset, MAE for MSG-CHN decreased from 699.6 to 653.4 compared to ProxyTTA-fast, and RMSE also slightly decreased (from 1120.3 to 1079.7), which may be due to the peculiarities of the error structure in certain areas of the images. However, on other datasets, such as SceneNet and ScanNet, CAPR shows consistently the same values for both MAE and RMSE (e.g., ScanNet MAE: 302.2 → 272.5, RMSE: 480.0 → 493.2).

Table 1.

Quality results for outdoor environments, from VOID to NYUv2, SceneNet and ScanNet

Method	**VOID→NYUv2**	**VOID→SceneNet**	**VOID→ScanNet**

	MAE	RMSE	MAE	RMSE	MAE	RMSE
MSG-CHN + Pretrained	1040.9	1528.9	281.2	645.0	687.9	1201.7
MSG-CHN + ProxyTTA(fast)	699.6	1120.3	192.7	424.4	302.2	480.0
MSG-CHN + CAPR (Ours)	653.4	1079.7	183.9	441.6	272.5	493.2
CostDCNet + Pretrained	189.1	446.7	173.3	443.2	144.3	458.6
CostDCNet + ProxyTTA(fast)	131.9	269.0	129.9	353.8	128.1	244.6
CostDCNet + ProxyTTA	95.8	203.8	125.7	357.1	68.1	162.3
CostDCNet + CAPRA (Ours)	86.1	174.4	121.5	336.5	62.9	154.6
NLSPN + Pretrained	388.9	702.8	167.3	438.7	233.3	431.2
NLSPN + ProxyTTA (fast)	168.4	309.4	124.6	357.5	104.0	232.8
NLSPN + ProxyTTA	124.4	240.7	113.9	333.4	74.8	166.6
NLSPN + CAPR (Ours)	111.1	210.8	108.9	276.4	70.7	151.3

The visual results analysis on the target dataset demonstrates the advantages of the proposed CAPR approach compared to previous adaptation methods. The structure of objects in the final depth predictions obtained by CAPR is restored more accurately in both homogeneous areas (e.g., smooth walls and boards) and complex boundary areas (e.g., furniture and textile edges) in all three scenes. The model's ability to avoid over-averaging and preserve sharp depth changes is particularly noticeable, as it is illustrated by less edge smoothing compared to baseline methods. Taken together, the results confirm that the introduction of cross-attention and bimodality in CAPR

improves the accuracy of depth recovery in new conditions, effectively overcoming the limitations of fixed projections and providing more robust adaptation in real-world.

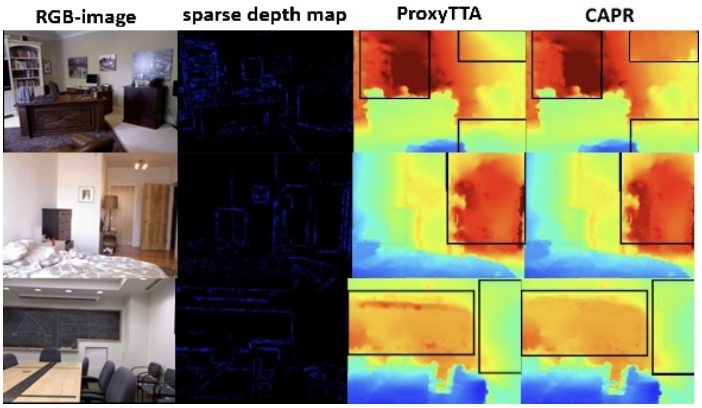

Figure 1. RGB image, sparse depth map, ProxyTTA dense depth map, CAPR dense depth map

Conclusion: The paper examined model depth restoration adaptation methods when the testing data distribution differs substantially from training data. Such changes occur frequently when scenes change (street to indoor) and within a single category (two indoor datasets). The standard practices of model transfer and BatchNorm adaptation layer updates fail to provide strong resistance against these types of distortions. The research developed CAPR (Cross-Attentive Proxy Refinement) as a solution to address these restrictions. CAPR stands apart from ProxyTTA through its dynamic cross-attention mechanism which adjusts image features based on scene and geometric information. The CAPR architecture supports mutual adaptation between modalities through its design which prevents feature collapse and strengthens intermodal connections. The Low-Rank Adaptation (LoRA) mechanism was implemented to decrease computational requirements and parametric expansion while preserving adaptation quality.

Although the method achieved success it remains limited by three factors: pre-training requirements for original datasets and higher memory usage than ProxyTTA and restricted adaptability to continuous domain transformations. Future research needs to address these identified limitations. Research should focus on enhancing

CAPR's ability to generalize across unfamiliar objects while adapting it for multiple computer vision applications including semantic segmentation and optical flow analysis and multimodal systems with unconventional sensors like IMU and thermal imager and radar. The proposed CAPR architecture presented in this paper represents an essential advancement toward depth regression models which adapt to scenes while being modality-aware and flexible. Looking forward, CAPR opens promising directions for real-time, multimodal, and continual adaptation in embodied AI, robotics, and autonomous driving applications.

References:

1. Cheng, Xinjing, Peng Wang, and Ruigang Yang. "Depth estimation via affinity learned with convolutional spatial propagation network." *Proceedings of the European conference on computer vision (ECCV)*. 2018.
2. Park, Jinsun, et al. "Non-local spatial propagation network for depth completion." *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*. Springer International Publishing, 2020.
3. Li, Ang, et al. "A multi-scale guided cascade hourglass network for depth completion." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020.
4. Kam, Jaewon, et al. "Costdcnet: Cost volume based depth completion for a single rgb-d image." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.
5. Park, Hyoungseob, Anjali Gupta, and Alex Wong. "Test-time adaptation for depth completion." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
6. Park, Hyoungseob, Anjali Gupta, and Alex Wong. "Test-time adaptation for depth completion." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

7. Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *ICLR* 1.2 (2022): 3.