Simultaneous Multi-Level Descriptor Learning and Semantic Segmentation for Domain-Specific Relocalization

Xiaolong Wu*, Yiye Chen*, Cédric Pradalier, and Patricio A. Vela

***Abstract*— This paper presents a semi-supervised framework for multi-level description learning aiming for robust and accurate camera relocalization across large perception variations. Our proposed network, namely DLSSNet, simultaneously learns weakly-supervised semantic segmentation and local feature description in the hierarchy. Therefore, the augmented descriptors, trained in an end-to-end manner, provide a more stable high-level representation for local feature dis-ambiguity. To facilitate end-to-end semantic description learning, the descriptor segmentation module is proposed to jointly learn semantic descriptors and cluster centers using standard semantic segmentation loss. We show that our model can be easily fine-tuned for domain-specific usage without any further semantic annotations, instead, requiring only 2D-2D pixel correspondences. Our learned descriptors, trained with our proposed pipeline, can significantly reduce the number of mismatches and thus boost the localization performance, which outperforms state-of-the-art descriptors or localization systems on the cross-season localization dataset.**

I. INTRODUCTION

Long-term visual localization is the problem of estimating the camera pose of a query image relative to 3D scene structure across appearance variations due to changes in time, weather, or seasons [1]. It is an important enabling module for robotic application domains involving long-term deployments in outdoor environments and others with highly variable visual conditions [2]–[5].

Classical approaches to localization rely on the local, image feature descriptors to establish 2D-3D correspondences for pose estimation [6]–[8]. Continued use of these approaches requires robust local descriptors sensitive to structural differences but not to changes in visual conditions. However, traditional feature descriptors are typically optimized for use when the query and database images are taken under similar visual conditions, as the emphasis is on robustness to the viewpoint [9], [10]. Mismatch errors increase if the localization and mapping stages occur under different visual circumstances. In contrast, data-driven approaches to feature detection and description are capable of establishing methods robust to both disturbance sources [11]–[14]. Metric learning techniques have contributed to these outcomes [15], since a deep neural network training process generates keypoint representations adapted to the data. Unfortunately, generalizing beyond the data remains problematic.

Another means to relax the sensitivity while preserving discriminatory power for point plus feature descriptor solutions has been to implement semantics-aided localization strategies. The introduction of high-level semantic information supplements the low-level descriptor information, resulting in fewer mismatches and more robust visual localization. This multi-level integration is beneficial, however, the increasing complexity of the network creates more annotation demands and requires attention with regards to how the layers should be trained due to the coupling. Consequently, the influence of network structure and learning methods on performance has not been fully explored. The efficacy of high-level guidance on top of multi-level description learning is unclear, meaning that evidence-supported guidance concerning network design for localization is nonexistent. Additionally, task-specific fine-tuning of semantics or semantics-augmented description learning has trouble addressing annotation needs or indicating how to reduce human-annotation demands.

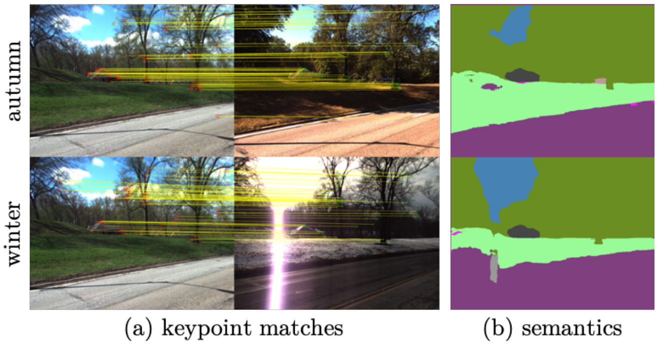

Fig. 1: **Matches and semantics from DLSSNet.** DLSSNet extracts keypoints with multi-level descriptors for image matching across significant appearance variations due to seasonal changes. Semantic segmentation is a side output.

We fill this gap by evaluating multi-level descriptors w or w/o high-level guidance in the context of long-term localization. Based on the results, we choose the best-performed network presented in this paper. In particular, we show that the semantic segmentation and local feature description can be simultaneously learnt within the hierarchy of a single deep network pipeline. A descriptor segmentation module jointly learns semantic descriptors and cluster means using a standard segmentation loss. The augmented multi-level descriptors, trained in an end-to-end manner, provide a more stable high-level representation for local feature disambiguation as compared to other architectures.

So that training complexity and data annotation needs are not limiting factors, we provide a pre-training process approach using standard semantic segmentation datasets, after which the network may be fine-tuned using only 2D-2D matches. The two-part process permits the domain-specific adaptation of the learnt localization module for addressing the challenges of long-term localization. Modifications have been made to facilitate efficient task-specific fine-tuning without extra semantic annotations. The learnt descriptors reduce the number of mismatches and thus boost localization performance. These outcomes are quantified using cross-season localization benchmarks and shown to outperform contemporary state-of-the-art baselines.

II. RELATED WORKS

A. Deep Feature Learning

Given that the classical paradigm for visual localization involved sequential feature detection followed by feature description, deep learning *detect-then-describe* pipelines naturally arose. Reflecting a similar sequential nature, detector networks first generate score maps for keypoint selection [16]–[18]. Next, small image patches centered at keypoint locations are cropped and input to descriptor networks to generate keypoint descriptions [15], [19]–[26]. The networks are independent of each other and trained that way. Unified architectures describe sequentially applied networks that can be jointly trained in an end-to-end manner [27], [28]. Though simplifying the training process, these unified architectures preserve network complexity. Importantly, they continue to process descriptors at the image patch level, which prevents learning contextual cues outside of the local patches.

For modern deep networks with high parametric degrees of freedom, it is sensible that both detection and description could be encoded within the same network and jointly learnt. This joint design describes a *detect-and-describe* deep feature learning approach that operates on full-sized images as opposed to patches [11]–[14]. Using a single encoding element improves computational efficiency due to shared weights. What distinguishes the methods are the self-supervised learning tactics and detector design. SuperPoint [11] uses a dual branch decoder with a heatmap keypoint detector on one of the decoder branches. R2D2 [13] has branches for repeatability and reliability maps that together recover keypoints of interest. The idea is to identify points that will almost always be detected and whose descriptions will be discriminative. Theoretically, image-wide convolutional structures permit learning of high-level contextual information, however contextual learning is limited by the stencil sizes. The learnt features are observed to still have fairly localized representations of image structure.

D2-Net [12] avoids a separate detection branch and builds a sequential process to perform keypoint detection directly from the pixel descriptors. It is more accurately called a *describe-to-detect* process since the descriptor outputs influence the detector. Detections based on feature vector outputs should induce higher-level feature learning that is less localized in nature. If so, it comes at the price of reduced keypoint localization accuracy. ASLFeat [14] instead modifies the *detect-and-describe* process by outputting multiple keypoint detector heatmaps at different layers within the network to recover low-, mid-, and high-level structure. Their fusion acts like a multi-scale keypoint detection process. Additional elements in the network provide robustness to image deformation arising from viewpoint changes. What is ultimately captured at the ”high” level will still be constrained by the stencils across the layers.

B. Semantics-Aided Localization

One form of higher-level contextual information is semantic scene knowledge. Semantic labels act as a weak supervisory signal to distinguish between correct and incorrect correspondences. A common strategy is to incorporate semantic labels into the matching stage of localization, where each 2D-3D match is assigned to a semantic consistency score [29], [30] to influence the RANSAC sampling for robust pose estimation. The bottleneck of such methods is the number of available classes, as the quantity directly translates to the discriminative power of potential matches. Self-supervised semantic learning methods that generate fine-grained segmentation networks overcome this problem [31].

Joint, multi-level descriptors combine the complementary information from semantics and geometric local features during the feature learning process. One group of them [32], [33] directly combines the features from off-the-shell networks trained on low- or high-level tasks for multi-level keypoint description. Although the unified frameworks have been delicately designed for end-to-end training, the fine-tuning of high-level representation still faces an efficacy problem due to multi-network multi-task learning needs [32]. Alternatively, multi-level descriptors concatenated from different stages of the network improve object detection, semantic segmentation, and part labeling [34]. Following this work, a hierarchical metric learning pipeline is proposed [35], and improved [36], to make the multi-level descriptors trainable end-to-end using simple correspondence contrastive loss. Other methods try to learn the multi-level descriptors in conjunction with auxiliary high-level tasks: the pixel-wise descriptors are extracted and concatenated from multiple layers of image retrieval network for localization tasks using merely image-level supervision [37]. Later, point-wise supervision [8] is incorporated along with the image correspondences for fine-grained point description learning, but using single-level descriptions.

III. LEARNING MULTI-LEVEL DESCRIPTION AND SEMANTIC SEGMENTATION

As illustrated in Fig. 2, our proposed Descriptor Learning and Semantic Segmentation Network (*DLSSNet*) is built upon PSPNet [41] pretained on semantic segmentation tasks. The feature embeddings from different stages of the network are first fused into two integrated descriptors (§III-A). Then, the resultant low- and high-level descriptors are utilized to accomplish three complementary tasks: low-level *detection-and-description* (§III-B), high-level semantic segmentation (§III-C), and multi-level description (§III-D).

Our network design takes advantage of the inherent hierarchy of CNNs to learn semantic and local geometric features from different layers of a single deep network [34]. Local geometric features identify low-level structures, such as corners, edges, etc., for keypoint localization and description. Contextual features, learned from semantic segmentation

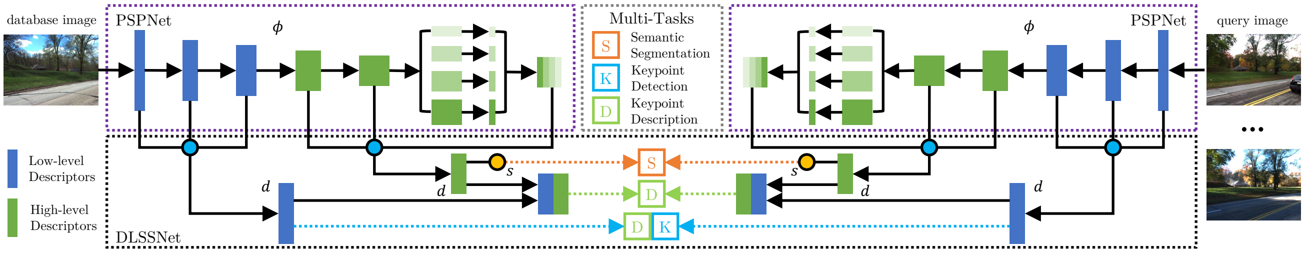

Fig. 2: **DLSSNet architecture** Our proposed network consumes a single image as input, which simultaneously predicts low-level keypoint detection and description, high-level semantic segmentation, and multi-level description within a single network.

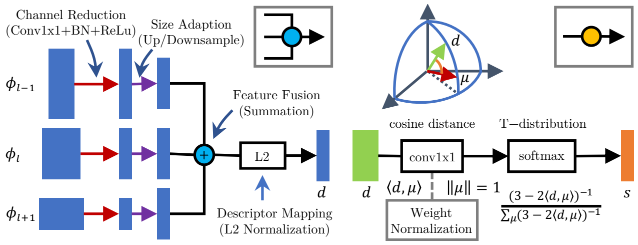
Channel Reduction (Conv1x1+BN+ReLu)
Size Adaption (Up/Downsample)
ϕ_{l-1}
ϕ_l
ϕ_{l+1}
Feature Fusion (Summation)
L2
d
Descriptor Mapping (L2 Normalization)
d
μ
cosine distance
T−distribution
d
conv1x1
softmax
s
$\langle d, \mu \rangle$
$\|\mu\| = 1$
Weight Normalization
$\frac{(3-2\langle d,\mu\rangle)^{-1}}{\sum_\mu (3-2\langle d,\mu\rangle)^{-1}}$

Fig. 3: **Sub-modules in DLSSNet**: multi-level feature fusion (left) and descriptor segmentation (right).

tasks, encode the high-level regional information into point description. The multi-level concatenated descriptors formulate a richer visual representation for keypoint matching.

A. *Multi-Level Feature Fusion*

The main task of our proposed feature fusion module is to establish a low-level geometric local descriptor and a high-level semantic descriptor by fusing feature embeddings from multiple network layers. We first divide deep features extracted from the bottlenecks (5 layers) and pyramid pooling module (1 layer) of PSPNet in a tuple format as shown in Fig. 2. Then, the two groups of feature embeddings are fed into our proposed feature fusion module for descriptor formulation. We group feature in a tuple format based on the observations that (1) state-of-the-art feature learning networks either truncate deep models after the third blocks [12] or utilize very shallow networks [11], [13], [14] for low-level high-resolution local description learning, and (2) simple fusion of lower-level features into high-level ones tend to be less effective for semantic segmentation [39].

As illustrated in Fig. 3 (left), our proposed multi-level feature fusion module takes the multi-level feature embeddings $\{\phi_{l-1}, \phi_l, \phi_{l+1}\}$ with different resolutions and channels as inputs and outputs a fused descriptor $\mathbf{d}$ for subsequent tasks. Channel reduction and size adaption modules first resize the feature maps to equal channels and same spatial resolution through individual *conv1x1+bn+relu* computations and *bilinear interpolation*, respectively. For computational efficiency, we implement feature summation, instead of concatenation, to fuse the generated features into a tensor of 128 channels at each resolution. The integrated descriptors are finally mapped into Spherical space for dense description formulation through channel-wise *L2-normalization*. The fused descriptors capture either geometrical information from bottom network layers or semantic cues from top ones.

B. *Learning Low-Level Detection-and-Description*

The low-level *description-and-detection* module here builds on D2-Net [12] to jointly learn a descriptor and keypoint detector using ground-truth point matches from SfM models. Considering the low-level natures of keypoint detection, *detection-and-description* uses low-level descriptors that not only maintain spatial resolution but also learn local details for accurate keypoint localization.

1) Detection: Given the low-level dense descriptor $\mathbf{d} \in \mathbb{R}^{H \times W \times C}$, the local spatial and channel-wise score maps, α and $\beta \in \mathbb{R}^{H \times W \times C}$, are computed for soft keypoint detection:

$$\alpha_{ij}^k = \frac{exp(d_{ij}^k)}{\sum_{(i',j') \in \mathscr{N}(i,j)} exp(d_{i'j'}^k)}, \quad \beta_{ij}^k = \frac{exp(d_{ij}^k)}{\max_{k'} exp(d_{ij}^{k'})} \qquad (1)$$

where $\mathscr{N}(i,j)$ is a set of 9 neighboring pixels around coordinate (i,j). Finally, the saliency score map $\mathbf{s} \in \mathbb{R}^{H \times W}$ weights the pixel-wise description loss after image-wise *L1-normalization*:

$$s_{ij} = \gamma_{ij} \Big/ \sum_{(i',j')} \gamma_{i'j'}, \text{ where } \gamma_{ij} = \max_k \alpha_{ij}^k \beta_{ij}^k \qquad (2)$$

2) Learning Detection-and-Description: Given a set of correspondences $\mathscr{C}$ from an image pair (I, I'), the *detection-and-description* loss of D2-Net [12], maximizes description distinctiveness at most repeatable keypoint locations:

$$\mathscr{L}_{dl}(I, I') = \frac{1}{|\mathscr{C}|} \sum_{c \in \mathscr{C}} \frac{s_c s_c'}{\sum_{k \in \mathscr{C}} s_k s_k'} \mathscr{M}(\mathbf{d}, \mathbf{d}') \qquad (3)$$

where s and s' are detection score maps (2) calculated from image I and I', respectively. $\mathscr{M}(\cdot,\cdot)$ is the ranking loss for description learning given a pair of descriptors $\mathbf{d}$ and $\mathbf{d}'$ of point correspondence.

Following the suggestion from ASLFeat [14], we utilize the hardest-contrastive loss [40] instead of hardest triplet loss in D2-Net [12] for better convergence:

$$\mathscr{M}(\mathbf{d}, \mathbf{d}') = \mathscr{M}(\mathbf{d}, \mathbf{d}')_+ + \mathscr{M}(\mathbf{d}, \mathbf{d}')_- \qquad (4)$$

where

$$\mathscr{M}(\mathbf{d}, \mathbf{d}')_+ = \max(0, D(\mathbf{d}, \mathbf{d}') - m_p)$$
$$\mathscr{M}(\mathbf{d}, \mathbf{d}')_- = \max(0, m_n - \min(\min_{\bar{\mathbf{d}} \neq \mathbf{d}} D(\bar{\mathbf{d}}, \mathbf{d}'), \min_{\bar{\mathbf{d}}' \neq \mathbf{d}'} D(\mathbf{d}, \bar{\mathbf{d}}')))$$

for $D(\cdot,\cdot)$ the distance between two descriptors. m_p and m_n are the positive and negative margins.

C. Learning High-Level Description Segmentation

Semantic segmentation, as an auxiliary task, is trained at deeper layers of the network to learn the high-level representation. The emphasis is on formulating and stabilizing the semantic descriptors for domain-specific applications. As such, no special effort is made to improve semantic segmentation performance. The uniqueness of our training is that the descriptor segmentation module jointly learns semantic descriptors $\mathbf{d}$ and cluster means μ in an end-to-end manner, where the probability distribution of semantic assignments is computed directly from learnt descriptors. It can be easily trained using point correspondence loss and supervised or weakly-supervised semantic segmentation loss.

1) Descriptor segmentation: Inspired by the recent advance of Deep Embedded Clustering (DEC) [42], a descriptor segmentation module is proposed to jointly learn the high-level descriptors $\mathbf{d} \in H \times W \times C$ and cluster centers $\mu \in C \times K$. It is achieved by mapping the descriptors from the spherical space to a lower-dimensional feature space that iteratively optimizes the semantic assignments $s \in H \times W \times K$.

Specifically, we use the t-distribution as a kernel to measure the similarity s between descriptor $\mathbf{d}$ and k^{th} cluster centroid μ_k at (i, j) coordinate:

$$s_{ij,k} = \frac{(1+\|\mathbf{d}_{ij,k}-\mu_k\|^2)^{-1}}{\sum_{k'}(1+\|\mathbf{d}_{ij,k'}-\mu_{k'}\|^2)^{-1}} = \frac{(3-2\mathbf{d}_{ij,k}^T\mu_k)^{-1}}{\sum_{k'}(3-2\mathbf{d}_{ij,k'}^T\mu_{k'})^{-1}} \tag{5}$$

As such, the semantic assignment s can be efficiently computed using standard convolutional operations *conv1x1+softmax* as shown in Fig. 3 (right), where the column weights of *conv1x1*, after *L2-normalization*, serve as cluster centers μ for joint optimization.

2) Learning Semantic Segmentation: Though DEC is designed for unsupervised learning, we advocate for supervised segmentation since it improves descriptor learning, as long as the ground-truth labels are available, due to the stability induced by supervision. To enable task-specific fine-tuning without additional data annotations, a weakly-supervised semantic segmentation pipeline is introduced to achieve efficient domain adaption. The losses for supervised and weakly-supervised training are detailed here.

The standard cross-entropy loss is chosen to supervise the model training using ground-truth semantic labels:

$$\mathscr{L}_{ce}(I) = \frac{1}{|I|}\sum_{(i,j)\in I} l_{ij}^T \log(s_{ij}) \tag{6}$$

where s stands for probability distribution of semantic assignments s computed from our proposed descriptor segmentation, and l is the ground-truth semantic labels.

For the cases that the semantic annotation is unavailable, the 2D-2D correspondences from SfM reconstructions can be used to enforce labeling consistency between acquired images as long as one of the images is taken at a similar condition as pre-trained datasets. As such, the ground-truth labels can be substituted by one-hot pseudo-label $\hat{l}$ calculated from the confidence of semantic assignment s as:

$$\mathscr{L}_{pce}(I,I') = \frac{1}{|\mathscr{C}|}\sum_{c\in\mathscr{C}} \hat{l}_c^T \log(s_c) \tag{7}$$

The pseudo-labels should be calculated from database images in localization datasets captured in favorable conditions [43].

3) Learning Point Correspondence: The priority of the descriptor segmentation module is to learn point-wise semantic descriptors for robust matching, where the semantic segmentation loss only enforces region-to-region similarity. To encourage a fine-grained high-level description, the point correspondence loss are incorporated for fine-grained semantic representation learning:

$$\mathscr{L}_{pc}(I,I') = \frac{1}{|\mathscr{C}|}\sum_{c\in\mathscr{C}} \mathscr{M}(d_c, d_c')_+ \,. \tag{8}$$

D. Learning Multi-Level Description

There are several options for merging the low-level geometric and high-level semantic descriptors. One is to combine two descriptors through element-wise *summation* and *L2-normalization* like in the feature fusion module and in [32]. In the presence of the semantic gap between two types of features, it is questionable that summation is capable of preserving the complementary information due to loss of separability. Feature concatenation is a better option at the expense of increased computational cost at the matching stage [34]–[38]. Having the keypoint detection scores (2) and the concanated descriptors, we train our multi-level descriptors in an end-to-end manner using the *detection-and-description*, hardest-contrastive loss in (3).

E. Training and Fine-Tuning

Fig. 4 depicts the training process, with semantic segmentation pre-training, followed by task-specific fine-tuning with weak supervision from 2D-2D correspondences. The two tasks iteratively optimize the description. To facillitate both training and fine-tuning, the combined loss is implemented:

$$\mathscr{L} = \lambda^l \mathscr{L}_{dl}^l + \lambda^h(\mathscr{L}_{(p)ce}^h + \mathscr{L}_{pc}^h) + \mathscr{L}_{dl}^m \tag{9}$$

where the superscripts l, h, and m refer to low-, high-, and multi-level tasks, and $\lambda^{(\cdot)}$ are weighting terms.

DLSSNet is pre-trained using semantic segmentation datasets in conjunction with a self-supervised synthetic image pairs. The ground-truth point correspondences are computed along with the randomly transformed image and semantic labels. Pre-training helps to initialize all three components of DLSSNet. A two-step pre-training procedure for descriptor learning initializes cluster centers without significantly altering the high-level representations. First, we fix the weights of PSPNet and optimize the introduced modules to convergence. Second, the whole network is jointly optimized.

Using domain adaptation ideas [43], we implement a weakly-supervised training pipeline to alleviate the demand for manually created annotations for long-term localization. The 2D-2D point correspondence from SfM enforces semantic and geometric consistency for domain-specific point descriptor fine-tuning. The pipeline takes advantage of the

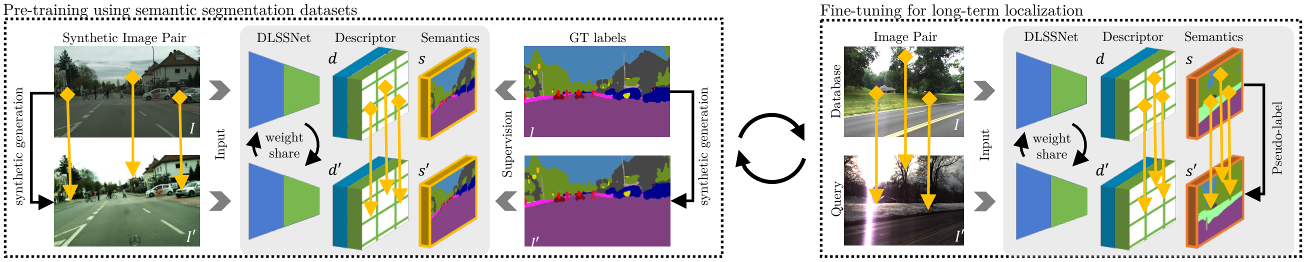

Fig. 4: **Training pipeline** of DLSSNet can be expressed as a two-stage procedure, where a supervised pre-training and a weakly-supervised task-specific fine-tuning are interleaved for semantics-stabilized description learning in the context of long-term localization.

TABLE I: Dataset Information

Dataset	Condition	Training Seq	Testing Seq	#Training Pairs	#Testing Pairs
Cityscapes [44]	urban	all	-	5,000	-
KITTI [45]	urban	05-08	-	10,000	-
RobotCar [46]	urban	all	-	6,511	-
CMU [47]	urban	6-8	2-5	28,766	46,569
	suburban	9-13	14-17		
	park	21-25	18-20		

observation that the database images in long-term localization datasets are usually captured during favorable conditions. It inherently enables knowledge transfer from easy to hard. To this end, the cross-entropy loss is optimized with pseudo-labels inferred from semantic assignments of database images. Implementing an interleaved training procedure maintains a stable semantic representation. It prevents trivial solutions during optimization of the high-level representations [43].

IV. Experimental Evaluation

In this section, the datasets used and implementation details are first described. Then, we investigate the effects of key parameters on the design of DLSSNet. Finally, we benchmark DLSSNet on the task of cross-season localization.

A. Datasets

DLSSNet is trained from image pairs gathered from three different sources. A semantic segmentation dataset (Cityscapes [44]) with synthetically generated pairs initialize the descriptors and clusters. Later, a visual odometry dataset (KITTI [45]) uses sequential images as pairs to refine the network, especially for local feature learning. Finally, the cross-season localization datasets (RobotCar-Seasons [46] and Extended CMU-Seasons [47]) with ground-truth cross-seasonal point correspondence (Cross-Seasons Correspondence Dataset [43]) fine-tune the whole network for cross-season localization. Table I summarizes the dataset information for training and testing. The Extended CMU-Seasons datasets splitting is based on the availability of corrrespondences for training, where the slices with ground-truth correspondence serves as training data and the rest as testing data.

B. Implementation Details

This section gives implementation details about training that most affect performance, and about evaluation.

1) Training Details: As a starting point, the PSPNet [41] (ResNet-101), pre-trained on Cityscapes [44] dataset, are selected for cross-season point description learning. To obtain dense ground-truth correspondence, image sequences are first fed into an SfM pipeline for dense reconstruction. Ground-truth correspondences are computed by estimating either 2D-3D or 3D-3D matches between image pairs. The 2D-3D matching strategy is usually applied to scenes with minor appearance changes, where standard 2D features will work. The 3D-3D method is used for long-term localization datasets, since 2D features are unreliable under large appearance changes. The ground-truth semantic annotations are either provided by semantic segmentation datasets or directly inferred from the network outputs as pseudo-labels in a weakly supervised pipeline.

As a preprocessing step, image pairs are first standardized to zero-mean unit-norm tensors, then augmented using random photometric transformations, such as brightness, contrast, and color noise. During the optimization, the loss described in (9) is computed from pairs with at least 50 point matches, and with loss balance factors $\lambda^l = \lambda^h = 0.4$. The SGD optimizer is used with a learning rate of 0.1. DLSSNet is trained on semantic segmentation tasks, CityScapes, with synthetic pairs for 30 epochs, where the synthetic data is simple and fast for rendering. Later, it is fine-tuned on KITTI, RobotCar-Seasons, and Extended CMU-Seasons for 50 epochs each for cross-season specific description learning.

2) Localization pipeline: Evaluation of cross-season localization uses a two-step pipeline. First, dense SfM models are generated from COLMAP [48] using the Multi-View Stereo (MVS) [49] pipeline as a default setting. Then, the query images are registered to 3D maps (image + depth map) using customized features. The 2D-3D matches, defined as mutual nearest neighbors, are used to realize camera poses using n-point-pose solver [50] inside a RANSAC loop [51]. To compare different deep features we use: (1) provided ground-truth poses to generate image pairs; and (2) SIFT [52] features to reconstruct dense depth maps using COLMAP with MVS option for all reference images.

3) Evaluation protocal: Camera localization is evaluated using camera pose recall, that is, the percentages of successfully localized images using the coarse-to-fine error tolerances $(0.5m, 2deg)/(1m, 5deg)/(5m, 10deg)$ for the ablation study and performance evaluation.

TABLE II: Comparison with State-of-the-Arts

Deep Descriptor	Extended CMU-Seasons					
	w/o SSMC [29]			w. SSMC [29]		
	urban	suburban	park	urban	suburban	park
SuperPoint [11]	89.8 / 91.2 / 92.5	85.9 / 89.7 / 91.5	75.9 / 81.3 / 87.1	91.4 / 92.8 / 95.5	88.3 / 91.2 / 93.2	81.9 / 83.1 / 90.0
D2-Net [12]	91.3 / 94.5 / 96.5	90.5 / 92.8 / 95.5	84.6 / 88.4 / 91.5	91.3 / 94.6 / 96.8	90.8 / 93.0 / 95.7	84.8 / 89.2 / 91.9
R2D2 [13]	92.6 / 93.7 / 95.3	90.8 / 93.2 / 95.2	81.7 / 87.3 / 91.9	93.1 / 94.6 / 96.5	91.4 / 92.8 / 95.7	85.8 / 89.1 / 91.7
ASLFeat [14]	92.7 / 94.4 / 96.5	90.9 / 93.9 / 95.8	85.3 / 88.9 / 91.6	**93.2** / 94.9 / 96.6	91.4 / 93.2 / 95.9	86.1 / 90.5 / 92.1
HML* [35]	92.4 / 94.6 / 96.2	91.2 / 93.7 / 95.4	85.3 / 88.8 / 91.2	93.0 / 94.8 / 96.5	91.6 / 92.8 / 95.7	86.8 / 90.4 / 92.1
SAND* [36]	92.9 / 94.6 / 96.7	91.3 / 93.7 / 95.8	85.1 / 89.0 / 91.5	93.1 / 94.7 / 96.7	91.6 / 92.9 / 96.0	86.9 / 90.7 / 92.3
DLSSNet	**93.2 / 95.4 / 97.0**	**92.4 / 94.2 / 96.2**	**88.4 / 91.6 / 92.5**	**93.2 / 95.4 / 97.0**	**92.4 / 94.2 / 96.2**	**88.4 / 91.6 / 92.5**

C. Ablation Study

To understand network design properties, we study the impact of: (1) different training datasets; (2) multi-level description fusion; (3) semantic description; and (4) high-level guidance from semantic segmentation using Extended CMU-Seasons dataset. Table III lists the ablation study outcomes.

Unlike state-of-the-art networks for large-scale web image retrieval, DLSSNet is trained using autonomous driving datasets for applicability of the semantic annotations. The table first lists percentages for successfully localized images when training on subsets of the training corpus. Synthetically generated image pairs don't provide satisfying performance (row 1). Incorporating real image pairs from KITTI sequences enables significant improvements at all scenes, while the fine-tuning at RobotCar-Seasons and Extended CMU-Seasons give another boost especially for park scenes (rows 2 & 3).

A major claim of this work is that multi-level descriptors combine the complementary strength of low- and high-level features for better long-term localization. The second part of the ablative study analyzes this proposal by comparing the single- and multi-level descriptors. The evaluation reveals that shallower features provide more accurate camera pose estimation than deeper ones (rows 4 & 5), while the high-level component, although less discriminative, provides sufficient complementary information to boost performance (row 6).

A third claim is that semantic segmentation, as high-level guidance, helps to learn better descriptors for camera relocalization. We compare the features learned w & w/o the auxiliary semantic segmentation task (rows 7 & 8). High-level descriptors learned from human-defined semantic segmentation tasks do provide better performance. Compared to semantic confidence of pre-defined labels, semantic descriptors learns a richer description for robust matching (rows 9 & 10). Similar to the finely-grained segmentation [31], the learned semantic descriptor enforces a finer data association during keypoint matching.

TABLE III: Ablative Study

Training Configuration	Extended CMU-Seasons		
	Urban	Suburban	park
Cityscapes	75.5 / 84.2 / 91.7	68.6 / 75.2 / 86.5	62.8 / 69.2 / 76.7
+KITTI	85.3 / 92.5 / 94.5	84.5 / 91.2 / 94.3	73.4 / 82.7 / 88.3
+RC/CMU	93.2 / 95.4 / 97.0	92.4 / 94.2 / 96.2	88.4 / 91.6 / 92.5
Low-lvl	92.7 / 94.6 / 96.5	89.3 / 92.9 / 95.1	83.3 / 86.9 / 89.6
High-lvl	65.3 / 72.2 / 86.3	64.6 / 73.9 / 83.3	60.2 / 68.2 / 75.7
Multi-lvl	93.2 / 95.4 / 97.0	92.4 / 94.2 / 96.2	88.4 / 91.6 / 92.5
w/o SemSeg	92.9 / 94.6 / 96.2	91.2 / 92.7 / 95.4	85.3 / 88.8 / 91.2
w SemSeg	93.2 / 95.4 / 97.0	92.4 / 94.2 / 96.2	88.4 / 91.6 / 92.5
SemSeg label	93.0 / 94.7 / 96.7	91.9 / 92.4 / 95.0	86.3 / 90.7 / 91.3
SemSeg desc	93.2 / 95.4 / 97.0	92.4 / 94.2 / 96.2	88.4 / 91.6 / 92.5

D. Comparison with State-of-Arts

DLSSNet is benchmarked on three scenes of the Extended CMU-Seasons dataset. As suggested[1], baselines include deep local features [11]–[14], as well as ”hypercolumns” [35], [36]. Also included is the option of a RANSAC-based, semantics-aided matching approach Semantic Segmentation Match Consistency (SSMC) [29].

Table II presents the results. As can be seen, DLSSNet outperforms state-of-the-art descriptors w or w/o the SSMC option, while being insensitive to the option itself. It validates the assertion the coupling semantics with local shape descriptor learning can help to build a richer representation for accurate camera relocalization. Of the tests, the *park* scene is the hardest. The vegetation inherently exhibits more appearance changes across seasons compared to the *urban* and *suburban* scenes. The improvement for *park* scenes further indicates the importance of the high-level description learning, which is the major contribution of our approach.

V. CONCLUSION

We described a single-network design, DLSSNet, with explicit semantic and geometric feature learning branches that extracted feature information from across multiple layers of the network and fused them according to best design principles. The intent is to explicitly enable the learning of multi-level (high and low) representations from the training imagery. A decoupled then coupled, multi-stage training process was described that lowers human annotation demands yet still enables coupled learning of the multi-level descriptors. The resulting network achieved strong performance on cross-season localization.

The city-centered semantic data sets used significantly constrain the training corpus. Therefore, all datasets used here concentrate on autonomous driving applications either within or across seasons. Other day-night and cross-weather localization datasets that do not match the setting cannot be applied, which is an opportunity for further study.

[1] https://www.visuallocalization.net

References

[1] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic *et al.*, "Benchmarking 6dof outdoor visual localization in changing conditions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8601–8610.

[2] F. Dayoub and T. Duckett, "An adaptive appearance-based map for long-term topological localization of mobile robots," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2008, pp. 3364–3369.

[3] T. Krajník, J. P. Fentanes, O. M. Mozos, T. Duckett, J. Ekekrantz, and M. Hanheide, "Long-term topological localisation for service robots in dynamic environments using spectral maps," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2014, pp. 4537–4542.

[4] C. Valgren and A. J. Lilienthal, "Sift, surf & seasons: Appearance-based long-term localization in outdoor environments," *Robotics and Autonomous Systems*, vol. 58, no. 2, pp. 149–156, 2010.

[5] E. Stenborg, C. Toft, and L. Hammarstrand, "Long-term visual localization using semantically segmented images," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 6484–6490.

[6] L. Liu, H. Li, and Y. Dai, "Efficient global 2d-3d matching for camera localization in a large-scale 3d map," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2372–2381.

[7] M. Geppert, P. Liu, Z. Cui, M. Pollefeys, and T. Sattler, "Efficient 2d-3d matching for multi-camera visual localization," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2019, pp. 5972–5978.

[8] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 716–12 725.

[9] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, "Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5173–5182.

[10] J. L. Schonberger, H. Hardmeier, T. Sattler, and M. Pollefeys, "Comparative evaluation of hand-crafted and learned local features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1482–1491.

[11] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018, pp. 224–236.

[12] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable cnn for joint detection and description of local features," *arXiv preprint arXiv:1905.03561*, 2019.

[13] J. Revaud, P. Weinzaepfel, C. De Souza, N. Pion, G. Csurka, Y. Cabon, and M. Humenberger, "R2d2: Repeatable and reliable detector and descriptor," *arXiv preprint arXiv:1906.06195*, 2019.

[14] Z. Luo, L. Zhou, X. Bai, H. Chen, J. Zhang, Y. Yao, S. Li, T. Fang, and L. Quan, "Aslfeat: Learning local features of accurate shape and localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6589–6598.

[15] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas, "Working hard to know your neighbor's margins: Local descriptor learning loss," in *Advances in Neural Information Processing Systems*, 2017, pp. 4826–4837.

[16] Y. Verdie, K. Yi, P. Fua, and V. Lepetit, "Tilde: A temporally invariant learned detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5279–5288.

[17] N. Savinov, A. Seki, L. Ladicky, T. Sattler, and M. Pollefeys, "Quad-networks: unsupervised learning to rank for interest point detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1822–1830.

[18] A. Barroso-Laguna, E. Riba, D. Ponsa, and K. Mikolajczyk, "Key.net: Keypoint detection by handcrafted and learned cnn filters," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 5836–5844.

[19] Y. Tian, B. Fan, and F. Wu, "L2-net: Deep learning of discriminative patch descriptor in euclidean space," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 661–669.

[20] K. He, Y. Lu, and S. Sclaroff, "Local descriptors optimized for average precision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 596–605.

[21] M. Keller, Z. Chen, F. Maffra, P. Schmuck, and M. Chli, "Learning deep descriptors with scale-aware triplet networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2762–2770.

[22] Z. Luo, T. Shen, L. Zhou, S. Zhu, R. Zhang, Y. Yao, T. Fang, and L. Quan, "Geodesc: Learning local descriptors by integrating geometry constraints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 168–183.

[23] D. Mishkin, F. Radenovic, and J. Matas, "Repeatability is not enough: Learning affine regions via discriminability," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 284–300.

[24] A. Mukundan, G. Tolias, and O. Chum, "Explicit spatial encoding for deep local descriptors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9394–9403.

[25] P. Ebel, A. Mishchuk, K. M. Yi, P. Fua, and E. Trulls, "Beyond cartesian representations for local descriptors," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 253–262.

[26] Y. Tian, X. Yu, B. Fan, F. Wu, H. Heijnen, and V. Balntas, "Sosnet: Second order similarity regularization for local descriptor learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 016–11 025.

[27] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 467–483.

[28] Y. Ono, E. Trulls, P. Fua, and K. M. Yi, "Lf-net: learning local features from images," in *Advances in Neural Information Processing Systems*, 2018, pp. 6234–6244.

[29] C. Toft, E. Stenborg, L. Hammarstrand, L. Brynte, M. Pollefeys, T. Sattler, and F. Kahl, "Semantic match consistency for long-term visual localization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 383–399.

[30] T. Shi, S. Shen, X. Gao, and L. Zhu, "Visual localization using sparse semantic 3d map," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 315–319.

[31] M. Larsson, E. Stenborg, C. Toft, L. Hammarstrand, T. Sattler, and F. Kahl, "Fine-grained segmentation networks: Self-supervised segmentation for improved long-term visual localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 31–41.

[32] Z. Luo, T. Shen, L. Zhou, J. Zhang, Y. Yao, S. Li, T. Fang, and L. Quan, "Contextdesc: Local descriptor augmentation with cross-modality context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2527–2536.

[33] S. Hong, K. Li, Y. Zhang, Z. Fu, M. Liu, and Y. Guo, "Learning local features with context aggregation for visual localization," *arXiv preprint arXiv:2005.12880*, 2020.

[34] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 447–456.

[35] M. E. Fathy, Q.-H. Tran, M. Zeeshan Zia, P. Vernaza, and M. Chandraker, "Hierarchical metric learning and matching for 2d and 3d geometric correspondences," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 803–819.

[36] J. Spencer, R. Bowden, and S. Hadfield, "Scale-adaptive neural dense features: Learning via hierarchical context aggregation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6200–6209.

[37] H. Germain, G. Bourmaud, and V. Lepetit, "Sparse-to-dense hypercolumn matching for long-term visual localization," in *Proceedings of the IEEE International Conference on 3D Vision (3DV)*, 2019, pp. 513–523.

[38] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[39] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, "Exfuse: Enhancing feature fusion for semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 269–284.

[40] C. Choy, J. Park, and V. Koltun, “Fully convolutional geometric features,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 8958–8966.

[41] J. Xie, R. Girshick, and A. Farhadi, “Unsupervised deep embedding for clustering analysis,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2016, pp. 478–487.

[42] M. Larsson, E. Stenborg, L. Hammarstrand, M. Pollefeys, T. Sattler, and F. Kahl, “A cross-season correspondence dataset for robust semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9532–9542.

[43] J. Min, J. Lee, J. Ponce, and M. Cho, “Hyperpixel flow: Semantic correspondence with multi-layer neural features,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 3395–3404.

[44] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223.

[45] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354–3361.

[46] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, “1 Year, 1000km: The Oxford RobotCar Dataset,” *The International Journal of Robotics Research (IJRR)*, vol. 36, no. 1, pp. 3–15, 2017.

[47] H. Badino, D. Huber, and T. Kanade, “The CMU Visual Localization Data Set,” http://3dvis.ri.cmu.edu/data-sets/localization, 2011.

[48] J. L. Schönberger and J.-M. Frahm, “Structure-from-motion revisited,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[49] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, “Pixelwise view selection for unstructured multi-view stereo,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[50] L. Kneip, D. Scaramuzza, and R. Siegwart, “A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 2969–2976.

[51] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[52] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.