

Semantic SuperPoint: A Deep Semantic Descriptor

Gabriel Soares Gama
 São Carlos School of Engineering
 University of São Paulo
 São Carlos, SP, Brazil
 gabriel_gama@usp.br

Nicolas dos Santos Rosa
 São Carlos School of Engineering
 University of São Paulo
 São Carlos, SP, Brazil
 nicolas.rosa@usp.br

Valdir Grassi Jr.
 São Carlos School of Engineering
 University of São Paulo
 São Carlos, SP, Brazil
 vgrassi@usp.br

Abstract—Several SLAM methods benefit from the use of semantic information. Most integrate photometric methods with high-level semantics such as object detection and semantic segmentation. We propose that adding a semantic segmentation decoder in a shared encoder architecture would help the descriptor decoder learn semantic information, improving the feature extractor. This would be a more robust approach than only using high-level semantic information since it would be intrinsically learned in the descriptor and would not depend on the final quality of the semantic prediction. To add this information, we take advantage of multi-task learning methods to improve accuracy and balance the performance of each task. The proposed models are evaluated according to detection and matching metrics on the HPatches dataset. The results show that the Semantic SuperPoint model performs better than the baseline one.

Index Terms—Computer Vision, Visual Odometry

I. INTRODUCTION

Visual Simultaneous Localization and Mapping (vSLAM) is the process of creating a map based on the visual information without a global reference, being used in multiple applications where GPS signal is not available or precise enough. In the context of vSLAM based on features, this is done by using the relative pose obtained by detecting interest points, generating a descriptor for them and matching the extracted keypoint in the previous frame to the current one. Those matches are then processed using algorithms similar to the normalized eight-point algorithm [3]. On top of that, loop closing and global optimization processes are normally implemented to improve the estimated trajectory, removing drift error.

All of that is dependent on the correct association of keypoints. Most feature extraction methods generate descriptors based solely on photometric information like ORB [24], SURF [2] and SIFT [19], with only the first one being both fast and reliable. Although they achieve excellent results in SLAM systems, handmade descriptors are highly susceptible to illumination changes, which decreases the quality of the pose estimation. This problem is important to be considered since

This work was supported by the São Paulo Research Foundation - FAPESP (grant number 2021/08117-0 and 2014/50851-0), by the Brazilian National Council for Scientific and Technological Development (grant number 465755/2014-3), and by the Coordination of Improvement of Higher Education Personnel - Brazil (Finance Code 001).

Code available at: <https://github.com/Gabriel-SGama/Semantic-SuperPoint>

978-1-6654-6280-8/22/\$31.00 ©2022 IEEE

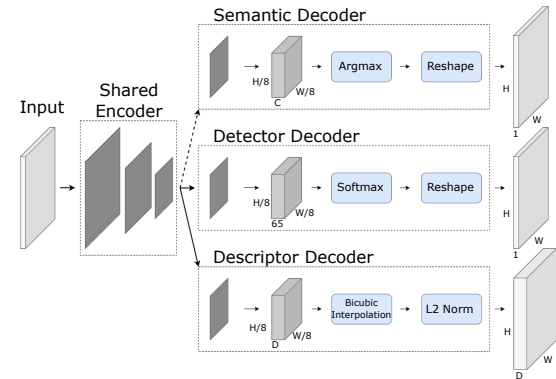


Fig. 1: **Semantic SuperPoint Model**. The dashed line represents that the semantic decoder was only used in training and C is the number of classes.

most SLAM applications are made to run in an uncontrolled environment.

Convolutional neural networks are normally used in computer vision tasks to increase robustness, and most often obtain better results than handmade methods [18], [26]. Similarly, methods such as SuperPoint [6] and LIFT [27] were developed to improve the coherency of feature extraction algorithms. Both are superior to ORB and present comparable or better results than SIFT while keeping a real-time performance.

In the context of deep learning, models that learn multiple related tasks can have a better result if trained with a shared encoder due to the inductive bias, as seen in the related literature [14], [21], [25]. In comparison to using multiple models and soft parameter sharing, the computational cost of using a shared encoder for multi-task is relatively low.

This article proposes adding semantic information intrinsically to the keypoint and descriptor, which would help to detect meaningful pixels and create a better descriptor. In this way, the pose estimation method can consider this semantic information as an additional criterion in the matching process of the extracted features. To accomplish that, multi-task learning methods are essential to effectively train the model. A diagram of the proposed model can be seen in Figure 1.

In this work, we also evaluate the models trained with and without the semantic decoder on the HPatches [1] according

to the metrics defined by DeTone et al. [6] and on the KITTI dataset [8] at the SLAM task using the ORB-SLAM2 [20] as the base framework. We show that the proposed Semantic SuperPoint (SSp) model improves the matching score metric and estimates better trajectories. For a 480×640 image, our model runs at **208 FPS** or **4.8 ms** per frame on an NVIDIA GTX 3090.

In summary, the key contributions of this paper are:

- Evaluation of multi-task learning methods applied to SuperPoint-like models.
- A novel training paradigm for deep feature extraction.
- A slight improvement in the Matching Score metric.

II. RELATED WORK

Semantic SLAM. Adding high-level semantic information to SLAM-related applications was previously reported in the literature on multiple occasions to improve methods based only on the intensities of the pixels. Examples of those projects are the use of high-level semantic information for loop closure detection with LiDAR 3D points to avoid local minimums [16], detection of dynamic objects to calculate their estimated speed and take that into account for a more accurate pose estimation process [10], segmentation of objects for better description considering a database of selected artifacts [11], and for a handmade semantic descriptor combined with a photometric one [7].

All of those highly depend on the quality of the network prediction, therefore are sensible to reflections and unknown objects. Even though the same can be said about our model, the semantic information only helps to extract better features and is not used in the inference, so it is more robust to misclassification.

Multi-task Learning. The simplest way to address multiple tasks is to do a uniform sum of each loss, as presented in Equation (1), in the specific case of $w_t = 1$ for every task t . This usually obtains the worst result because of the different nature of the tasks and scales of the loss functions used to minimize them. Consequently, the model can converge prioritizing one task over another.

One way to avoid this is to compute a weighted sum of each loss according to a positive scalar w_t . This has a better result than the uniform sum. However, it does not achieve the same result when compared to multi-task learning methods, requiring a grid search for the weighting parameters, which is time-consuming and computationally costly. The mentioned approach tries to minimize the following loss objective:

$$L_{total} = \sum_{t=1}^N w_t L_t, \quad (1)$$

where w_t is the weight that scales the t -th task loss L_t , and N is the number of tasks.

To avoid this undesired convergence behavior and the grid search, this work implements two multi-task learning methods proposed in [14], [21].

The first method [14] is based on the uncertainty modeling to weigh losses and derive a multi-task loss function by maximizing a Gaussian likelihood with homoscedastic uncertainty to classification and regression problems.

Let $f^{\mathbf{W}}(\mathbf{x})$ be the output of a neural network with \mathbf{W} weights and input \mathbf{x} and y for the respective label. For regression problems, considering a normal distribution, the log-likelihood associated is:

$$\log p(y|f^{\mathbf{W}}(\mathbf{x})) \propto -\frac{1}{2\sigma^2} \|y - f^{\mathbf{W}}(\mathbf{x})\|^2 - \log \sigma, \quad (2)$$

where σ is an observation noise scalar. In the case of a classification problem:

$$\log (y = c|f^{\mathbf{W}}(\mathbf{x}), \sigma) \propto -\frac{1}{\sigma^2} f^{\mathbf{W}}(\mathbf{x}) - \log \sigma, \quad (3)$$

where c is the predicted class.

The second approach [21] consists in finding a common descent direction for the shared encoder. This is done by computing the gradient related to each task, normalizing them, and searching for weighting coefficients that minimize the norm of the convex combination, obtaining the central direction.

To consider the gradient history, a period of T iterations of the gradients' norm of each task are tracked and, in case of some task diverging, the central dir. is pulled toward it as the tensioner's idea. Nakamura et al. [21] shows that by doing this the shared encoder parameters are adjusted to equally benefit all tasks, besides decreasing the convergence time significantly.

For a more detailed explanation, please refer to the original article of each method.

SuperPoint. SuperPoint [6] is a deep feature extractor trained by self-supervision that can perform in real-time while executing both keypoint detection and descriptor decoders.

For the detection decoder, a base detector MagicPoint is trained in the Synthetic Shapes dataset, which renders simple 2D objects, such as quadrilaterals, triangles, lines, and ellipses. In the dataset, the keypoint label is created by marking certain junctions positions, end of segments, and ellipses center.

The MagicPoint model is then used to extract the interest points in the selected dataset by performing a number N_h of homographic adaptations and combining its output to generate a final heatmap, filtered by Non-Maximum Suppression (NMS) [12].

The next step is joint training with both detector and descriptor decoders. A set with the original image and its warped version is passed through the model, outputting the detected interest points and the descriptors. Since the warp transformation is known, the same goes for the position relation between each pixel. Thus, pairs of matches and non-matches can be created and compared to the ground truth, illustrated in Figure 2.

III. SEMANTIC SUPERPOINT MODEL

SSp is a fully convolutional neural network that receives a grayscale image as input and outputs a heatmap, descriptor, and semantic segmentation prediction. The architecture uses

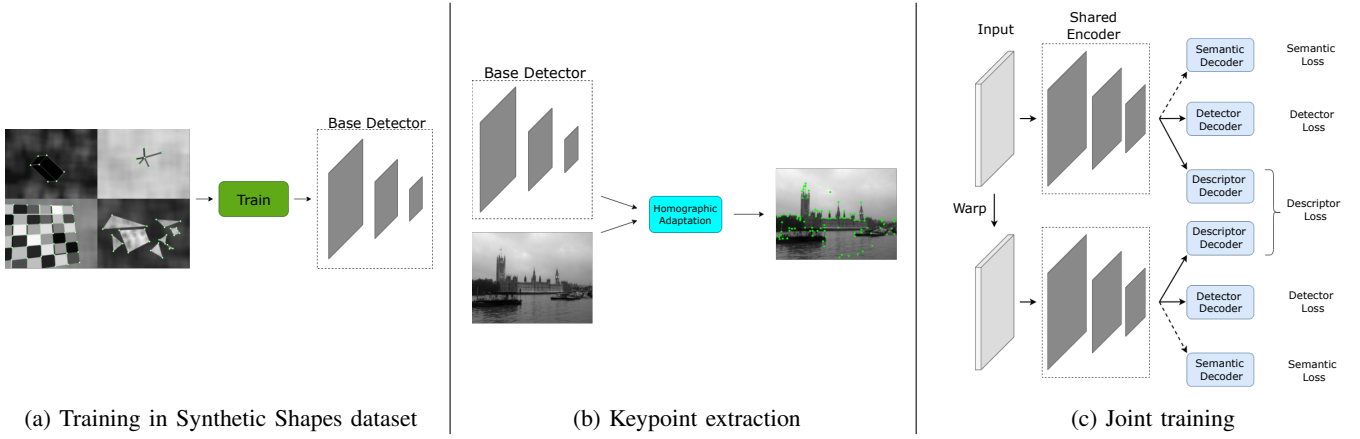


Fig. 2: **Semantic SuperPoint training overview.** A base betector is trained in the Synthetic Shapes dataset and used to extract pseudo-ground truth interest points in unseen real images using N_h homographic transformations. The joint training step uses the generated labels with the semantic label to train the SSp model. Adapted from DeTone et al. [6].

a single shared encoder and its output (intermediary features) passes through each decoder.

Due to the aggressive data augmentation process and the large variety of classes in the MS-COCO dataset [17], used for training, the semantic prediction is not usable in practical application. Nonetheless, adding a semantic decoder for training improved the quality of the deep feature extractor. It can therefore be said that rather than being a model, SSp is more like a training method.

A. Shared Encoder

According to the open source implementation of SuperPoint (Sp) [13], the encoder is based on the U-Net [23], using `double_conv` blocks, which consists of $2 \times (\text{conv}2d - \text{BN} - \text{ReLU})$ and three layers of max pooling to reduce the dimension between them. Therefore the output has a shape of $\frac{H}{8} \times \frac{W}{8} \times C_{enc}$. We adopt $C_{enc} = 256$ as the encoder output dimension. The same encoder model is used for Sp and SSp architectures.

B. Detector Decoder

Both the detector and descriptor decoders are the same as SuperPoint's original model. The detector has a shape of $\frac{H}{8} \times \frac{W}{8} \times 65$ to avoid upsampling operations and has one dummy dimension 'no keypoints'. The loss function used was the binary cross entropy (BCE):

$$L_d = \sum_{n=1}^N [y_n \log(\text{Softmax}(x_n)) + (1 - y_n) \log(1 - \text{Softmax}(x_n))], \quad (4)$$

where N is the batch size.

C. Descriptor Decoder

For the descriptor decoder, also to avoid adding computational cost related to the upsampling, the output is shaped as $\frac{H}{8} \times \frac{W}{8} \times D$ and to extract the descriptor, it passes

through a Bicubic Interpolation and is normalized. To train the descriptor, it is necessary to relate two images and evaluate the matches and non-matches while balancing both influences. This was done by dividing the loss value by the number of positive (n_p) and negative (n_n) matches in the Hinge loss:

$$L_{desc} = \underbrace{\frac{s}{n_p} \max(0, m_p - d^T d')}_{L_p} + \underbrace{\frac{(1-s)}{n_n} \max(0, d^T d' - m_n)}_{L_n}, \quad (5)$$

where $m_p = 1$, $m_n = 0.2$, d and d' are the descriptor decoder outputs for the normal and warped image, respectively.

D. Semantic Decoder

The semantic decoder consists of a `conv3x3-BN-ReLU` block followed by a `conv1x1`. The cross-entropy (CE) loss was used for training this decoder:

$$L_s = - \sum_{n=1}^N \sum_{c=1}^C w_c \log(\text{Softmax}(x_c)) y_{n,c}. \quad (6)$$

IV. MULTI-TASK LOSS

In this section, we discuss the losses applied in the training process. In all cases, the semantic loss will be included to avoid rewriting the equations, but the non-semantic loss can be obtained by removing the L_{s_i} terms.

A. Uniform loss

For comparison purposes, the model was also trained with the uniform sum loss function as follows:

$$L_{total} = L_{d_1} + L_{d_2} + \lambda L_{desc} + L_{s_1} + L_{s_2}. \quad (7)$$

The losses L_{d_i} and L_{s_i} are obtained from the interest point detector and the semantic decoder, respectively. The

subscribed index 1 is for the normal image and 2 for the warped version.

In the original SuperPoint article [6], they use the term λ to weight the descriptor loss. In our case, the code was build on the unofficial implementation of Jau et al. [13], and they use $\lambda = 1$. Besides, the model was able to learn both tasks without the need for tuning the value λ .

B. Uncertainty loss

In the case of the uncertainty loss proposed by Kendall et al. [14], the detection loss can be directly related to the classification case, Equation (3), since the BCE loss behaves the same way as a CE loss for 2 classes. The same goes for the semantic decoder as it also uses CE.

The descriptor loss is more complex. To jointly calculate the terms of the hinge loss, the assumption that the losses are not correlated must be true. Even though this is not the case for most multi-task models, the approximation remains valid. The problem with the hinge loss is that the positive and negative terms are inversely proportional. The more the positive loss improves, the negative loss will not get better at the same rate as the model would create a bias towards positive matches and vice versa.

The descriptor loss was considered as two independent regression problems, each one with a related normal distribution. Considering the loss given by Equation (5), it is expected that the output $D_{out} = f^{\mathbf{W}_{sh,desc}}(\mathbf{x}_1)^T f^{\mathbf{W}_{sh,desc}}(\mathbf{x}_2)$ is going to be grouped toward m_p or m_n for the positive matches ($s = 1$) and the negative matches ($s = 0$), respectively. So we can describe the distribution as:

$$p(\mathbf{y}|D_{out}, s) = s\mathcal{N}_p(D_{out}, \sigma_p^2) + (1 - s)\mathcal{N}_n(D_{out}, \sigma_n^2), \quad (8)$$

where \mathcal{N}_i and σ_i are the normal distribution and variance associated with each match type.

Expanding the normal distributions and approximating the loss terms, we obtain:

$$\log p(\mathbf{y}|D_{out}) \propto -\frac{1}{2\sigma^2} (L_p + L_n) - \log \sigma. \quad (9)$$

See Appendix VIII-A for more detail. The normal and warped loss terms are summed because they represent the same objective on the same weights, so the joint probability can be modeled as:

$$\begin{aligned} p(\mathbf{y}_{d_1}, \mathbf{y}_{d_2}, \mathbf{y}_{desc}, \mathbf{y}_{s_1}, \mathbf{y}_{s_2} | f^{\mathbf{W}}(\mathbf{x}_1), f^{\mathbf{W}}(\mathbf{x}_2)) = \\ p(y_{d_1}, y_{d_2} | f^{\mathbf{W}_{sh,d}}(\mathbf{x}_1), f^{\mathbf{W}_{sh,d}}(\mathbf{x}_2)) p(y_{desc} | D_{out}) \\ p(\mathbf{y}_{s_1} = c_1, \mathbf{y}_{s_2} = c_2 | f^{\mathbf{W}_{sh,s}}(\mathbf{x}_1), f^{\mathbf{W}_{sh,s}}(\mathbf{x}_2)). \end{aligned} \quad (10)$$

So the final equation for the uncertainty loss is given by:

$$\begin{aligned} -\log p(\mathbf{y}_{d_1}, \mathbf{y}_{d_2}, \mathbf{y}_{desc}, \mathbf{y}_{s_1}, \mathbf{y}_{s_2} | f^{\mathbf{W}}(\mathbf{x}_1), f^{\mathbf{W}}(\mathbf{x}_2)) \propto \\ (L_{d_1} + L_{d_2}) \exp(\eta_d) + \frac{L_{desc}}{2} \exp(\eta_{desc}) \\ + (L_{s_1} + L_{s_2}) \exp(\eta_s) + \eta_d + \frac{\eta_{desc}}{2} + \eta_s = L_{total}. \end{aligned} \quad (11)$$

In order to avoid numerical instability during training, we do the variable change: $\eta_i = 2 \log \sigma_i$. The initial values for $\eta_d, \eta_{desc}, \eta_s$ were 1.0, 2.0, 1.0, respectively.

C. Central dir. + tensor

To apply the Central dir. + tensor method the normal and warped loss were summed to simplify the common gradient descent calculation and the descriptor loss was treated as a single task.

The α term, which is used to regulate the tension factor's sensitivity, was set to 0.3, as indicated by the original authors [21].

V. DATA PREPARATION

The MS-COCO 2017 dataset consists of 123k 640x480 RGB images split into 118k images for training and 5k for validation, with 133 segmented classes. This large amount and variability of data can provide the model with the necessary generalization capability to perform in several conditions while being similar to the 2014 one, used to train the original SuperPoint model.

A. Keypoint extraction

In order to create the pseudo-ground truth labels for the keypoint extractor, a pretrained MagicPoint model in the Synthetic Shapes dataset from Jau et al. [13] was applied to the MS-COCO dataset. It was used $N_h = 100$ as the number of homograph transformations, resized to a resolution of 240×320 to decrease training time.

VI. TRAINING

All training and evaluation processes were done in a machine with an AMD Ryzen 9 5950X 16-Core Processor and an NVIDIA GTX 3090 24GB using PyTorch [22].

For data augmentation, multiple transformations, such as Gaussian noise, scale, rotation, translation, and others were applied to improve the network's robustness to illumination and viewpoint changes. The descriptor dimension is $D = 256$ for all experiments.

A. Sp & SSp training

The Sp and SSp models were trained using the Adam optimizer [15] for 200k iterations with batch size 16. We tested with a fixed learning rate as in the original article, and with a starting value of 0.0025 with polynomial decay and an end value of 0.001. The first option worked better for the non-semantic version and the second one was better for the SSp model.

We also varied the multi-task loss used for each model, as described in Section IV. Empirically, we found that the central dir. + tensor method took too long to converge. So we used the model saved at the 100Kth iteration trained by the uncertainty loss as the starting point, and trained for more 100K iterations using the central dir. + tensor optimization method with the same learning rate parameters.

In every variation, the models were saved every interval of 5K iterations and evaluated as described in Section VII.

To rank each model, we adopt the matching score as the main metric, since it is the percentage of inliers. Hence, it represents an overall performance of the keypoint detection and descriptor generation. Furthermore, it is the most relatable to SLAM methods.

VII. EVALUATION

A. HPatches

The HPatches dataset consists of 116 scenes each one with six different images. The first 57 scenes are subject to illumination changes and in the other 59 scenes, viewpoint changes.

The models were evaluated according to the metrics: repeatability (Rep.), localization error (MLE), homography estimation ($e = \{1, 3, 5\}$), nearest neighbor mean average precision (NN mAP), and matching score (M.S.). Each metric is described in detail by DeTone et al. [6].

Considering the abbreviations `unc` for the uncertainty loss, `ct` for central dir. + tensor and `uni` for the uniform loss, the results of each model are presented in Table I.

TABLE I: Evaluation on HPatches dataset

Feature extractor	Homography Estimation			Detector metrics		Descriptor metrics	
	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$	Rep.	MLE	NN mAP.	M.S.
ORB ^a	.150	.395	.538	.641	1.157	.735	.266
SIFT ^a	.424	.676	.759	.495	0.833	.694	.313
LIFT ^a	.284	.598	.717	.449	1.102	.664	.315
Sp + uni (baseline)	.476	.748	.817	.599	1.017	.864	.519
Sp + unc	.460	.745	.812	.599	1.010	.864	.520
Sp + ct	.493	.753	.812	.602	1.001	.862	.519
SSp + uni (ours)	.398	.710	.797	.584	1.052	.843	.506
SSp + unc (ours)	.450	.745	.816	.598	1.005	.864	.522
SSp + ct (ours)	.466	.762	.805	.598	0.999	.858	.519
Pretrained model [6]	.497	.766	.840	.610	1.086	.843	.540

^aOriginal results reported by DeTone et al. [6].

All metrics are higher is better, except MLE.

Analyzing the results of the SSp model, the uniform loss had the worst result compared to the others SuperPoint models. The uncertainty and central dir. methods improved significantly the initial result obtained using the uniform loss for the SSp model and had a very similar performance in the case of the Sp model.

Considering the matching score, the SSp model had the best result with a slight improvement of **+0.2%** over the baseline Sp. As for the Homography Estimation it had a better result than the Sp + unc for $\epsilon = 5$ while balancing the detector metrics.

Thus, adding semantic information can help deep feature extraction methods, but still needs some improvement to increase all metrics.

It was not possible to replicate the result obtained from the Magic Leaps pretrained model [6] since the official implementation is not public available as well its specifics. For this reason, the results were not directly compared with the pretrained model, instead we considered the Sp + uni as baseline for comparison.

B. KITTI

The best version of each model was evaluated in the KITTI dataset according to the implementation SuperPoint_SLAM [4]. Each sequence was executed 10 times and the Absolute Pose Error (APE) and Relative Pose Error (RPE) metrics were extracted with the open source lib `evo` [9]. The mean values and the standard variation together with the p-values are exhibited in Table II.

Note that the SLAM application used here is only for comparison purposes as the matching time adds to much computational cost, being responsible for most of it [5]. To deal with these issues, approaches similar to GCNv2 [5] can be used.

VIII. CONCLUSIONS

We have shown that adding semantic information in a shared encoder based architecture in combination with multi-task learning methods can improve deep feature extraction methods and, in this case, without increasing computational cost in the inference step.

Future works can tune the intensity of the data augmentation process and the complexity of the training dataset to balance generalization capability and semantic learning or train the feature extractor using another approach that does not rely on intensive data augmentation.

Additionally, if achieved a satisfying semantic prediction, it is possible to combine the high-level semantic information into the already existing semantic SLAM methods for further improvements.

APPENDIX

A. Descriptor distribution

The descriptor distribution is initially defined as two normal distributions as in Equation (8), copied bellow for convenience:

$$p(\mathbf{y}_{desc}|D_{out}, s) = s\mathcal{N}_p(D_{out}, \sigma_p^2) + (1-s)\mathcal{N}_n(D_{out}, \sigma_n^2). \quad (12)$$

Expanding the norm distribution we have:

$$p(\mathbf{y}_{desc}|D_{out}, s) = \frac{s}{\sigma_p\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\mathbf{y}_{desc} - D_{out}}{\sigma_p}\right)^2\right] + \frac{1-s}{\sigma_n\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\mathbf{y}_{desc} - D_{out}}{\sigma_n}\right)^2\right]. \quad (13)$$

Considering the same variance for both losses and substituting the term $(y - D)^2$ for the respective loss we have: [20]

$$p(\mathbf{y}_{desc}|D_{out}) = \frac{1}{\sigma\sqrt{2\pi}} \left[\exp\left(\frac{-L_p}{2\sigma^2}\right) + \exp\left(\frac{-L_n}{2\sigma^2}\right) \right]. \quad (14)$$

The s term is no longer needed because each loss is related to each match case. Since we are not balancing the losses between each other, we define the likelihood as:

$$p(\mathbf{y}_{desc}|D_{out}) \propto \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(L_p + L_n)}{2\sigma^2}\right). \quad (15)$$

TABLE II: Mean APE and RPE in the KITTI dataset for each model

Sequence	ATE			RPE		
	Sp	SSp	p-value	Sp	SSp	p-value
00	6.767 ± 1.112	6.651 ± 0.614	0.705	0.124 ± 0.028	0.130 ± 0.032	0.804
01	286.709 ± 140.875	209.985 ± 130.585	0.131	4.779 ± 1.910	3.991 ± 2.183	0.174
02	22.459 ± 2.428	22.314 ± 2.988	0.821	0.098 ± 0.008	0.098 ± 0.009	0.705
03	1.319 ± 0.142	1.630 ± 0.262	0.003	0.040 ± 0.002	0.046 ± 0.004	0.007
04	0.840 ± 0.135	0.905 ± 0.261	0.940	0.034 ± 0.004	0.039 ± 0.008	0.406
05	5.803 ± 1.233	6.315 ± 2.931	0.821	0.172 ± 0.046	0.199 ± 0.101	1.000
06	11.953 ± 0.939	11.833 ± 0.415	0.406	0.240 ± 0.048	0.267 ± 0.075	0.326
07	3.388 ± 2.947	2.124 ± 0.516	1.000	0.108 ± 0.071	0.084 ± 0.010	0.884
08	31.324 ± 1.736	26.700 ± 1.013	0.000	0.273 ± 0.012	0.235 ± 0.009	0.000
09	35.945 ± 2.140	31.788 ± 1.595	0.001	0.172 ± 0.013	0.164 ± 0.010	0.226
10	5.515 ± 0.437	4.953 ± 0.316	0.023	0.064 ± 0.003	0.060 ± 0.003	0.019

So the log likelihood is:

$$\log p(\mathbf{y}_{desc} | D_{out}) \propto -\frac{1}{2\sigma^2} (L_p + L_n) - \log \sigma. \quad (16)$$

REFERENCES

- Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *CVPR*, 2017.
- Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In Aleš Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision – ECCV 2006*, pages 404–417, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.
- W. Chojnacki and M.J. Brooks. Revisiting hartley’s normalized eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1172–1177, 2003.
- Chengqi Deng, Kaitao Qiu, Rong Xiong, and Chunlin Zhou. Comparative study of deep learning based features in slam. In *2019 4th Asia-Pacific Conference on Intelligent Robot Systems (ACIRS)*, pages 250–254. IEEE, 2019.
- Chengqi Deng, Kaitao Qiu, Rong Xiong, and Chunlin Zhou. Comparative study of deep learning based features in slam. In *2019 4th Asia-Pacific Conference on Intelligent Robot Systems (ACIRS)*, pages 250–254. IEEE, 2019.
- Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 337–33712, 2018.
- Baofu Fang, Gaofei Mei, Xiaohui Yuan, Le Wang, Zaijun Wang, and Junyang Wang. Visual slam for robot navigation in healthcare facility. *Pattern Recognition*, 113:107822, 2021.
- Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- Michael Grupp. evo: Python package for the evaluation of odometry and slam. <https://github.com/MichaelGrupp/evo>, 2017.
- Mina Henein, Jun Zhang, Robert Mahony, and Viorela Ila. Dynamic slam: The need for speed. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2123–2129. IEEE, 2020.
- K. Himri, P. Ridao, N. Gracias, A. Palomer, N. Palomeras, and R. Pi. Semantic slam for an auv using object recognition from point clouds. *IFAC-PapersOnLine*, 51(29):360–365, 2018. 11th IFAC Conference on Control Applications in Marine Systems, Robotics, and Vehicles CAMS 2018.
- Jan Hosang, Rodrigo Benenson, and Bernt Schiele. Learning non-maximum suppression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4507–4515, 2017.
- You-Yi Jau, Rui Zhu, Hao Su, and Manmohan Chandraker. Deep keypoint-based camera pose estimation with geometric constraints. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4950–4957. IEEE, 2020.
- Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *3rd International Conference for Learning Representations, San Diego, 2015*, 2014.
- Lin Li, Xin Kong, Xiangrui Zhao, Tianxin Huang, and Yong Liu. Semantic scan context: A novel semantic-based loop-closure method for lidar slam. *Auton. Robots*, 46(4):535–551, apr 2022.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014.
- Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot MultiBox detector. In *Computer Vision – ECCV 2016*, pages 21–37. Springer International Publishing, 2016.
- David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.
- Raul Mur-Artal and Juan D. Tardos. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-d cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, oct 2017.
- Angelica Tiemi Mizuno Nakamura, Valdir Grassi Jr, and Denis Fernando Wolf. Leveraging convergence behavior to balance conflicting tasks in multi-task learning. *Neurocomputing*, 2022.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571, 2011.
- Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 525–536. Curran Associates, Inc., 2018.
- Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European conference on computer vision*, pages 467–483. Springer, 2016.