2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

SNF-Feat: Semantic-Guided Negative-Sample-Free Representation Learning for Local Feature Extraction

Xun Zhou[1], Qingqing Yan[1], Minghao Zhu[1], Mengxian Hu[1], Chengju Liu[1,2] and Qijun Chen[1]

***Abstract*— Local feature extraction constitutes a foundational module crucial for numerous downstream tasks of computer vision. Its primary challenge lies in the generation of discriminative feature representations. Prior methodologies have employed contrastive learning within their pipelines, yet have encountered limitations stemming from inherent conflicts within their training data, including the ambiguity of negative samples and the distortion of positive samples. In this study, we propose a semantic-guided negative-sample-free method for local feature learning, denoted as *SNF-Feat*. Our framework entails dense patch-level representation learning without reliance on negative samples, aiming to ensure that descriptors derived from transformed views of the same local area exhibit predictive capability towards each other. To assess the impact of positive sample distortion, we harness high-level semantic information to derive point-wise loss weights. Furthermore, we establish a self-supervised feature learning paradigm that extends our utilization of datasets. Experimental results demonstrate the superior performance of our method across a range of typical datasets and tasks in comparison to *state-of-the-art* approaches.**

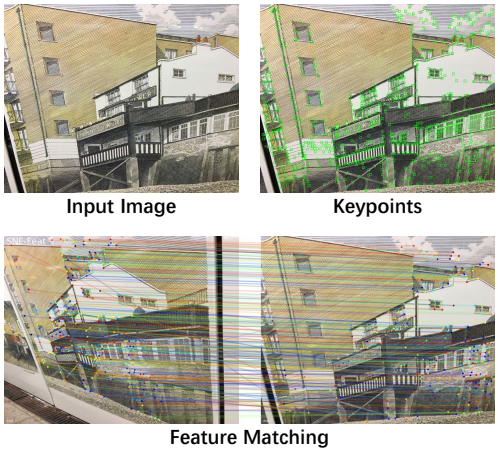

Fig. 1. **Features extracted by SNF-Feat.** Many of our features are situated within regions of high uniqueness and distinctiveness, thereby endowing our feature with superior matching capabilities.

I. Introduction

Local feature extraction serves as a foundational component for numerous downstream tasks in computer vision, such as visual localization, SLAM (Simultaneous Localization And Mapping), and SfM (Structure-from-Motion)[1][2][3]. A local feature encompasses both a keypoint's spatial location within the image and a descriptor, a representation vector encoding the local patch information surrounding the keypoint.[4]

Researchers have identified the "describe-then-detect" pipeline[5][6] as optimal for feature learning, wherein discriminative descriptors are first generated for all points, followed by the detection of keypoints exhibiting high discrimination and repeatability. Consequently, descriptor generation emerges as the primary challenge. We anticipate that identical descriptors characterize the same local areas across various images, while different descriptors distinguish disparate local areas. *State-of-the-art* (SOTA) methodologies often employ contrastive learning as their foundational approach[7], necessitating pairs of positive and negative samples. Nevertheless, their efficacy is hindered by inherent conflicts within the training data, including the ambiguity of negative sample pairs and the distortion of positive ones.

In contrastive-learning-based methods, descriptor learning necessitates triplet training samples. For instance, considering a feature point $\boldsymbol{p_1}$ in I_1 as the anchor point, along with the transformation information between I_1 and I_2, we can ascertain its corresponding point $\boldsymbol{p_{2+}}$ in I_2 as the positive point, and designate another point $\boldsymbol{p_{2-}}$ in I_2 as the negative point. For the triplet $(\boldsymbol{p_1}, \boldsymbol{p_{2+}}, \boldsymbol{p_{2-}})$, descriptors $(\boldsymbol{d_1}, \boldsymbol{d_{2+}}, \boldsymbol{d_{2-}})$ are derived from descriptor sets. To foster the acquisition of discriminative descriptors, these methodologies aim to minimize the distance between $\boldsymbol{d_1}$ and $\boldsymbol{d_{2+}}$, while simultaneously maximizing the distance between $\boldsymbol{d_1}$ and $\boldsymbol{d_{2-}}$. [8]

SOTA methodologies contend that negative samples are crucial to avert model collapse. Nevertheless, in typical scenes, local patches surrounding $\boldsymbol{p_1}$ and hand-crafted $\boldsymbol{p_{2-}}$ may exhibit a degree of similarity, such as certain cyclic structural components depicted on the left in Figure 2. This inherent ambiguity introduces false negatives, potentially leading to model degradation. Given the difficulty in establishing a definitive criterion for discerning entirely dissimilar point pairs, we naturally contemplate the omission of negative samples.

Recent studies indicate that negative samples are not indispensable for image-level representation learning. Grill et al.[9] and Chen et al.[10] have demonstrated that solely utilizing positive samples can forestall model collapse through certain training strategies. In this paper, we propose a dense patch-level representation learning approach devoid of negative samples for descriptor training. Our aim is that descriptors of image patches from transformed views of the same area can exhibit predictive capability towards each other, obviating the need for blind assignment of

[1]Department of Control Science and Engineering, Tongji University, Shanghai, China, [2]Shanghai Institute of Intelligent Science and Technology, Tongji University {zhouxun, qyan_0131, 2310202, humengxian, liuchengju, qjchen}@tongji.edu.cn

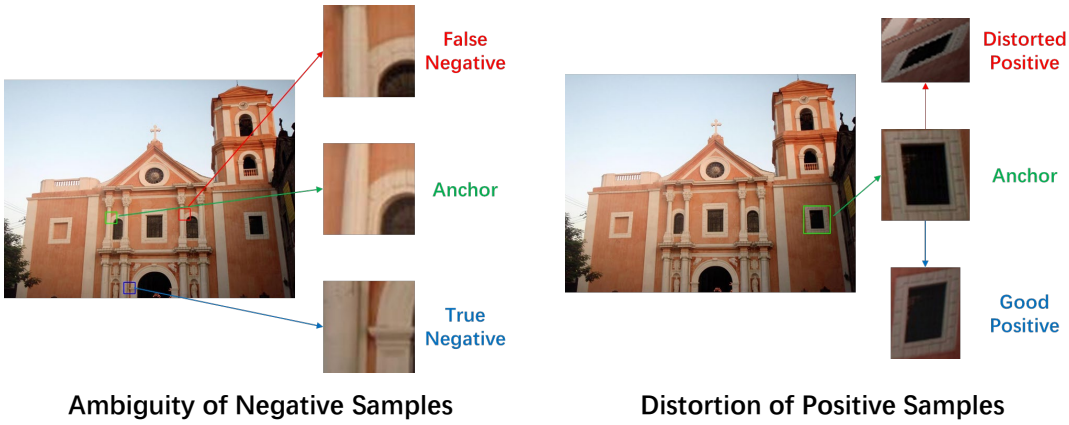

Fig. 2. **Illustration of negative ambiguity and positive distortion.** True negative samples exhibit markedly distinct appearances compared to the anchor sample, whereas false negatives, although differing in spatial location from the anchor, are challenging to differentiate from the anchor. Good positive samples maintain the structural integrity of the anchor, whereas distorted positive samples seldom exhibit meaningful structural coherence.

negative pairs. We employ a stop-gradient mechanism to mitigate model collapse. Experimental results demonstrate that our method achieves comparable accuracy and superior efficiency compared to methods that necessitate negative samples.

Another concern that has been insufficiently addressed is the distortion of positive samples. Representation learning mandates that positive sample pairs originate from widely varied viewpoints, environmental conditions, or self-supervised transformations. However, such random variations can potentially distort the scene structure, thereby diminishing positive confidence and potentially impairing optimization, as illustrated on the right in Figure 2. Consequently, we endeavor to assess the semantic consistency of synthetic image pairs using a real-time off-the-shelf semantic parsing model and incorporate weights into our prediction loss to mitigate the impact of distorted positive samples. Experimental findings demonstrate that semantic assistance enhances the performance of our negative-sample-free model.

SOTA methodologies traditionally rely on externally measured camera poses and scene depth maps to establish positive correspondences, yet the stringent requirement for precise ground truth imposes limitations on their dataset utilization capabilities. In this study, we employ random homography matrices to generate synthetic image pairs as positive samples, enabling us to utilize diverse image datasets and facilitate self-supervised feature learning, thereby enhancing the generalization ability of our model.

In summary, our contributions are as follow:

- We propose a negative-sample-free representation learning method aimed at extracting discriminative local features, thereby circumventing ambiguous point correspondences.
- We leverage high-level semantic information to construct consistency weights for the foundational loss, enabling the assessment of positive sample distortion for robust training.
- We establish a self-supervised feature learning framework applicable to widely-distributed datasets. Experimental results demonstrate the outstanding performance of our method across challenging vision tasks.

II. Related Work

A. Local Feature Learning

Early deep feature detectors initially mimicked handcrafted features, adhering to the detect-then-describe pipeline[11][12][13], which seemed intuitive. These detectors attempted to replace detection or/and description modules using Convolutional Neural Networks (CNNs). SuperPoint[3] devised a detect-and-describe pipeline where detection and description share the encoder while maintaining independent decoders. ALIKE[14] proposed a partly differentiable keypoint detection module.

Keypoints are primarily regarded as distinguishable points, prompting the adoption of the describe-then-detect pipeline. D2-Net[5] utilized a triplet loss for training the description module and generated a score map from descriptors in a non-learnable manner. ASLFeat[7] improved upon D2-net by incorporating deformable convolution layers to enhance keypoint localization. R2D2[15] introduced an additional indirect repeatable loss. Policy gradient techniques are used by DISK[16] to tackle the issue of discreteness encountered during the selection of sparse keypoints. CAPS[17] and PosFeat[6] designed an epipolar loss for weakly supervised descriptor learning. PUW-Feat[18] devised a progressive and unified pipeline for description and detection, aiming to enhance training robustness, and simplified keypoint learning.

In the describe-then-detect paradigm, keypoints learning modules can capture a wealth of high-level information, rendering them more accurate and robust, thus becoming the mainstream of local feature learning. However, despite their advantages, SOTA methods still rely on negative sample pairs, which may introduce harmful ambiguity.

B. Representation Learning

The feature descriptor, serving as the core task in our local feature learning method, can be viewed as a dense patch-

level instance of representation learning. Contrastive learning emerged as an early solution to representation learning, wherein it attracts positive sample pairs while repulsing negative sample pairs[19]. Notably, SimCLR[20] simply treats all other samples in the current batch as negative samples, thus requiring a large batch size for optimal performance. On the other hand, Moco[21] constructs a dynamic dictionary with a queue to store historical negative samples and transforms one branch into a momentum encoder to maintain inter-sample consistency.

Certain methods aim to avoid explicit utilization of negative samples. For instance, SwAV[22] simultaneously clusters all historical samples while enforcing consistency between cluster assignments from different views of the same image, with cluster centers implicitly playing the negative role.

More recently, researchers have discovered that negative samples can be completely disregarded in representation learning. BYOL[9] exclusively employs positive samples to train discriminative representations with the assistance of a momentum encoder. SimSiam[10] has demonstrated that a simple "stop-gradient" strategy can prevent model collapse in negative-sample-free learning. However, these approaches are predominantly focused on the image-level.

III. Proposed Method

In this part, we elaborate on our method, **SNF-Feat**. Our model constitutes a deep neural network that receives an RGB image I as input and produces both a dense descriptor map D and a keypoint score map S. We commence by presenting our network architecture and its utilization across various stages in Section III-A. The fundamental negative-sample-free training pipeline is elucidated in Section III-B. Furthermore, we introduce the semantic-guided loss for robust training in Section III-C. Lastly, Section III-D delves into the implementation specifics of our dataset and the settings of training parameters. The comprehensive training pipeline is illustrated in Figure 3.

A. Network Architecture

Our network comprises three modules: the encoder, projector $\mathcal{F}_{pj}$, and predictor $\mathcal{F}_{pd}$. Throughout the training process, all these modules are utilized; however, during deployment, only the encoder remains relevant. The encoder consists of two components: a descriptor encoder $\mathcal{F}_d$ and a keypoint encoder $\mathcal{F}_k$, which undergo training in two distinct stages. The descriptor encoder represents our primary challenge and is pivotal to our methodology.

During the descriptor training stage, when presented with a pair of images ($I_1 \in \mathbb{R}^{h_1 \times w_1 \times 3}, I_2 \in \mathbb{R}^{h_2 \times w_2 \times 3}$) depicting the same scene along with their corresponding points set $\mathcal{C}$, the description encoder generates dense representation maps

$$D_1 = L2\text{-}norm(\mathcal{F}_d(I_1) \in \mathbb{R}^{h_1 \times w_1 \times c_d}) \tag{1}$$

$$D_2 = L2\text{-}norm(\mathcal{F}_d(I_2) \in \mathbb{R}^{h_2 \times w_2 \times c_d}) \tag{2}$$

They can be interpreted as descriptors encapsulating information about the local patches surrounding each pixel within the images.

Subsequently, the projector $\mathcal{F}_{pj}$, implemented as a Multilayer Perceptron (MLP), is utilized to map the dense descriptors into a latent space

$$G_1 = \mathcal{F}_{pj}(D_1) \in \mathbb{R}^{h_1 \times w_1 \times c_g} \tag{3}$$

$$G_2 = \mathcal{F}_{pj}(D_2) \in \mathbb{R}^{h_2 \times w_2 \times c_g}. \tag{4}$$

Finally, we employ another MLP, known as the predictor $\mathcal{F}_{pd}$, to facilitate the mutual prediction of corresponding feature representations

$$Z_1 = \mathcal{F}_{pd}(G_1) \in \mathbb{R}^{h_1 \times w_1 \times c_z} \tag{5}$$

$$Z_2 = \mathcal{F}_{pd}(G_2) \in \mathbb{R}^{h_2 \times w_2 \times c_z} \tag{6}$$

During the detection training stage, we require pairs of images (I_1, I_2) along with the correspondence $\mathcal{C}$ as the input. Utilizing these inputs, we generate descriptor maps (D_1, D_2) through the description encoder. Subsequently, we employ the keypoint encoder $\mathcal{F}_k$, a shallow CNN branch, to produce keypoint score maps

$$S_1 = \mathcal{F}_k(D_1) \in \mathbb{R}^{h_1 \times w_1 \times 1} \tag{7}$$

$$S_2 = \mathcal{F}_k(D_2) \in \mathbb{R}^{h_2 \times w_2 \times 1} \tag{8}$$

which encapsulate the confidence levels associated with each point on the images, thereby serving as effective feature keypoints.

B. None-Negative Loss

Initially, let's focus on the descriptor training stage.

We commence with a single RGB image I_1 as input. Subsequently, a homography matrix H is randomly generated and applied to I_1 to yield a transformed image I_2, along with a dense correspondence set $\mathcal{C}$. Importantly, our approach is self-supervised, obviating the need for external ground truth such as poses or depth maps.

Let's assume we have a point $\boldsymbol{p_1^i}$ in I_1 as the anchor point. By referencing $\mathcal{C}$, we can ascertain its corresponding point $\boldsymbol{p_2^i}$ in I_2. These points represent the 2D projection of the same spatial point $\boldsymbol{p^i}$, albeit observed from different viewpoints and environmental conditions. Consequently, descriptors $\boldsymbol{d_1^i}$ and $\boldsymbol{d_2^i}$ are obtained for $\boldsymbol{p_1^i}$ and $\boldsymbol{p_2^i}$, respectively, from D_1 and D_2. Our objective in this paper is for descriptors $\boldsymbol{d_1^i}$ and $\boldsymbol{d_2^i}$ to predict each other, thereby eliminating the need for exhaustive mining of negative sample pairs to learn discriminative local feature representations.

The descriptor loss of a certain point $\boldsymbol{p^i}$ is

$$\begin{aligned}\mathcal{L}_d(\boldsymbol{p^i}) = 1 - \frac{1}{2}[&cosim(\boldsymbol{z_1^i}, stopgrad(\boldsymbol{g_2^i})) \\ &+ cosim(\boldsymbol{z_2^i}, stopgrad(\boldsymbol{g_1^i}))],\end{aligned} \tag{9}$$

Here we don't directly utilize the original descriptors themselves for mutual prediction. Instead, we manipulate the outputs of the projector and predictor modules $\boldsymbol{g_1^i} \subset G_1, \boldsymbol{g_2^i} \subset G_2$ and $\boldsymbol{z_1^i} \subset Z_1, \boldsymbol{z_2^i} \subset Z_2$. The cosine similarity metric,

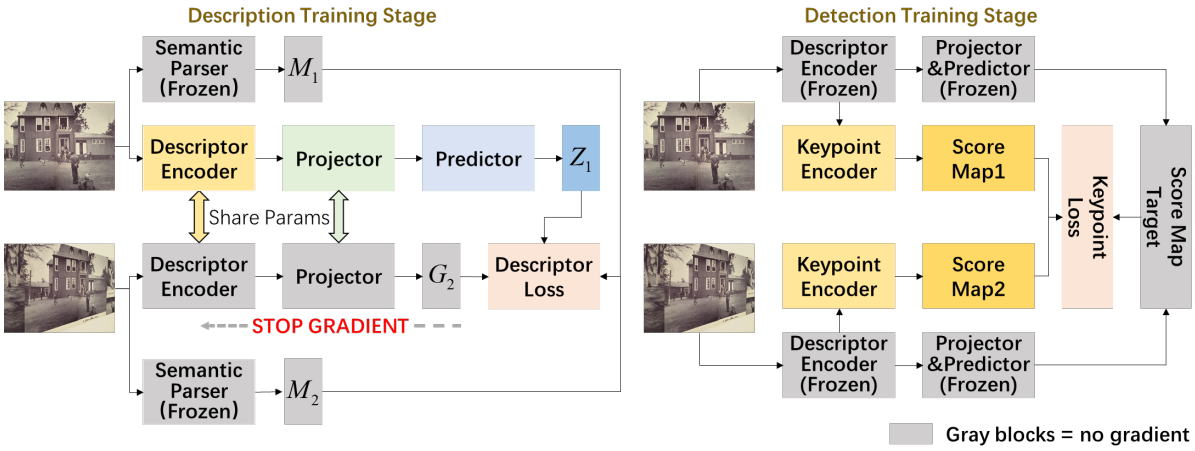

Fig. 3. **SNF-Feat Training Pipeline.** During the descriptor training phase, we employ a Siamese network architecture in conjunction with an auxiliary predictor, stop-gradient strategy, and semantic-assistance. We anticipate that corresponding descriptors will exhibit predictive capabilities towards one another. Our descriptor loss function is symmetric, albeit only half is depicted in the figure for enhanced clarity. In the detector training phase, we optimize the keypoint score maps based on the values derived from the descriptor loss.

denoted as *cosim*, serves as the evaluation criterion for mutual prediction accuracy. The primary strategy employed is *stopgrad*, which involves treating the object of this operation as a constant to halt the backward gradient flow[10], thereby ensuring the effective functioning of the model.

Subsequently, we can establish the fundamental structure of the descriptor loss:

$$\mathcal{L}_d = \frac{1}{|\mathcal{C}|} \sum_{\boldsymbol{p^i} \in \mathcal{C}} \mathcal{L}_d(\boldsymbol{p^i}). \tag{10}$$

During the initial stage, we exclusively train the descriptor encoder while disregarding the keypoint encoder. Once the description training converges, we freeze the descriptor encoder and proceed to train the keypoint encoder. As previously discussed, the descriptor loss value serves as an indicator of the matching capability of each point, with corresponding point pairs sharing identical loss values. Consequently, we can construct the training target for score maps using descriptor loss values. Subsequently, we optimize the keypoint encoder with respect to this target to extract discriminative and repeatable keypoints.

The keypoint score loss is

$$\mathcal{L}_k = \frac{1}{|\mathcal{C}|} \sum_{\boldsymbol{p^i} \in \mathcal{C}} \sqrt{(\frac{1}{2}(s_1^i + s_2^i) - (1 - \mathcal{L}_d(\boldsymbol{p^i})))^2} \tag{11}$$

where $s_1^i \subset S_1$ is the keypoint score of $\boldsymbol{p_1^i}$ in I_1, $s_2^i \subset S_2$ is the keypoint score of $\boldsymbol{p_2^i}$ in I_2 and $\mathcal{L}_d(\boldsymbol{p^i})$ is the descriptor loss value of the shared spatial point $\boldsymbol{p_i}$.

C. Semantic Consistency Weights

Various transformations between positive image pairs have the potential to distort the underlying scene structure, and in some cases, may even obliterate it, particularly with the random homography transformations utilized in this study. Put differently, the reliability of positive pairs cannot be strictly guaranteed, and such distortion may exacerbate performance degradation. To address this issue, we propose the incorporation of a semantic consistency metric as weights on the basic loss to assess the reliability of positive sample pairs.

To implement this, we introduce a pretrained off-the-shelf semantic parsing model $\mathcal{F}_{sm}$ with the same backbone architecture as our encoder. However, we exclude the final classification head layer and upsample the output to match the resolution of the input image. During each training iteration, the image pair is simultaneously processed with $\mathcal{F}_{sm}$ and the descriptor encoder.

$$M_1 = \mathcal{F}_{sm}(I_1) \in \mathbb{R}^{h_1 \times w_1 \times c_m} \tag{12}$$

$$M_2 = \mathcal{F}_{sm}(I_1) \in \mathbb{R}^{h_2 \times w_2 \times c_m} \tag{13}$$

Subsequently, we compute the point-wise Jensen-Shannon divergence[23] between M_1 and M_2 based on the the correspondence set $\mathcal{C}$:

$$w(\boldsymbol{p^i}) = 1 - JS(\boldsymbol{m_1^i} || \boldsymbol{m_2^i}). \tag{14}$$

where $\boldsymbol{m_1^i} \subset M_1$ and $\boldsymbol{m_2^i} \subset M_2$ are the semantic parsing outputs of the spatial point $\boldsymbol{p^i}$ in I_1 and I_2. A high $w(\boldsymbol{p^i})$ indicates that the semantic parsing model identifies the consistency of the inner structure between $\boldsymbol{p_1^i}$ and $\boldsymbol{p_2^i}$, thereby affirming the reliability of this positive point pair. It is worth noting that although the upper bound of the Jensen-Shannon divergence is $\ln 2$, we select 1 as the threshold for our semantic weights. This decision is made to constrain the diversity of weights, ensuring that the semantic model merely serves as an assisting component.

Finally we can build the complete format of our descriptor loss with semantic-guided weights

$$\hat{\mathcal{L}_d} = \sum_{\boldsymbol{p^i} \in \mathcal{C}} \frac{w(\boldsymbol{p^i})}{\sum\limits_{\boldsymbol{p^i} \in \mathcal{C}} w(\boldsymbol{p^i})} \mathcal{L}_d(\boldsymbol{p^i}). \tag{15}$$

Hence points exhibiting higher semantic consistency assume greater importance in loss computation.

It's important to note that the semantic weights are exclusively utilized during the description training stage. Once this stage converges, indicating that the description encoder has effectively acquired the ability to produce discriminative feature representations, we revert to employing the basic descriptor loss value as the target during the detection training stage.

D. Implementation Details

Network Details. For the crucial description encoder module, we implement a novel real-time network architecture called NDNet[25]. We omit its classification head layer and upsample the output to match the resolution of the input image, resulting in a dense descriptor map with $c_d = 256$. The semantic parsing model shares the same structure as the description encoder and $c_m = 256$, but is initialized with weights pretrained by a semantic segmentation task. The projector consists of a 3-layer MLP with $c_g = 128$. The first two layers incorporate batch normalization and ReLU activation functions, while the final layer includes only batch normalization. The predictor is constructed as a 2-layer MLP featuring a bottleneck structure with $c_z = 128$. Only the first layer includes batch normalization and ReLU activation functions. Finally, the detection encoder is comprised of a 3-layer shallow CNN. A Sigmoid activation function is applied to the output of the detection encoder to confine the keypoint scores within the range $[0, 1]$.

Training Details. We utilize Microsoft COCO[24] dataset, renowned for its wide distribution, to ensure the generalization ability of our model. Input images are randomly cropped to a uniform shape of 256×256. We implement a random homography matrix generation tool. Furthermore, we employ colorjitter and gaussion blur to simulate environment changes. We utilize the SGD optimizer with Nesterov momentum set at 0.9 and weight decay at 1e-4. The initial learning rate is set to 2e-4 and decreases exponentially at a rate of 0.9. Our model is trained on two NVIDIA RTX-4090 GPUs. During the description training stage, convergence is achieved after 150 epochs, while the detection training stage converges within 30 epochs.

Deployment. During deployment on real downstream tasks, only the description encoder and keypoint encoder are utilized. It's important to normalize descriptors using L2 normalization. Additionally, non-maximum suppression is applied to the keypoint score map, followed by the extraction of points with high scores as final feature points.

IV. Evaluation

We compare our method with SOTA methods, they can be categorized into three types:

- Detect-Then/And-Describe Methods: HesAffNet[26] with HardNet++[27], SIFT[4] with ContextDesc[13], SIFT with CAPS[17], SuperPoint[3]
- Describe-then-Detect Methods: R2D2[15], D2-Net[5], ASLFeat[7], DISK[16], and PosFeat [6]

TABLE I

MMAScore results from different methods on the Hpatches dataset.

Methods	MMAScore Overall	MMAScore Illumination	MMAScore Viewpoint
HAN + HN++	0.628	0.636	0.620
SIFT + ContextDesc	0.641	0.623	0.659
SIFT + CAPS	0.694	0.765	0.626
SuperPoint	0.660	0.725	0.597
R2D2	0.691	0.727	0.656
D2-Net	0.529	0.607	0.454
ASLFeat	0.736	0.791	0.683
DISK	0.771	0.820	0.725
PosFeat	0.736	0.819	0.657
SNF-Feat(Ours)	0.794	0.887	0.704

- Matcher methods: SuperGlue[28] + SuperPoint, Sparse-NCNet[29], LoFTR[30], Patch2Pix[31], CoAM[32]

For experiment results tables in this section, red means this result is the best result in this metric, green means the second best.

A. Feature Matching

In this experiment, we assess the performance of our method through a standard feature matching task conducted on the Hpatches[33] dataset. This dataset comprises 116 sequences, each consisting of 6 images with known homography matrices. The diversity within each sequence is categorized into two aspects: illumination and viewpoint variations. For every sequence, we designate the first image as the reference, matching the remaining images against it.

Our evaluation metric, mean matching accuracy *MMA*[34], quantifies the ratio of correct matches to possible matches. We vary the matching threshold Th_m from 1 pixel to 10 pixels, and compute a weighted sum of MMA across different thresholds[6] to derive an overall evaluation.

$$MMAScore = \frac{\sum_{Th_m=1}^{10} [(2 - 0.1Th_m) \cdot MMA@Th_m]}{\sum_{Th_m=1}^{10} (2 - 0.1Th_m)} \tag{16}$$

Results. As depicted in Figure 4 and summarized in Table I, our SNF-Feat achieves SOTA performance on the Hpatches dataset. When confronted with illumination changes, SNF-Feat surpasses all previous methods, showcasing its robustness to challenging environmental variations. In the face of viewpoint changes, only DISK[16] exhibits superior performance to our method. However, it's important to note that DISK relies on camera pose transformations and dense depth maps, which are pre-computed by a resource-intensive SfM tool to establish pixel-to-pixel correspondences. Additionally, DISK employs exhaustive reinforcement learning for training keypoint locations. In contrast, our method does not require external supervision or reinforcement learning, making it simpler to train and transfer to other datasets, thereby offering superior deployment performance.

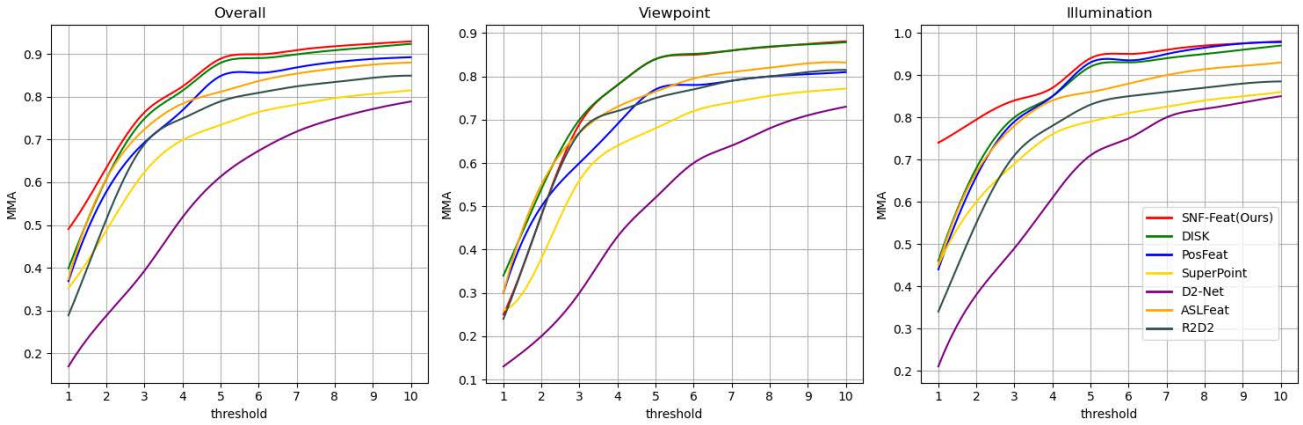

Fig. 4. **MMA** results of different methods on various matching thresholds in the Hpatches dataset with viewpoint and illumination change.

TABLE II
ABLATION EXPERIMENTS OF SNF-FEAT

Negative Sampling Strategy	Semantic Guiding	MMAScore	Mem/GB
random	✗	0.665	23.8
hardest	✗	0.776	29.9
hardest	✓	0.787	39.8
none	✗	0.762	16.8
none	✓	0.794	26.7

TABLE III
VISUAL LOCALIZATION RESULTS OF DIFFERENT METHODS ON THE AACHEN-DAY-NIGHT DATASET.

Methods	$(0.5m, 2\circ)$	$(1m, 5\circ)$	$(5m, 10\circ)$
SuperPoint	73.9	78.1	90.5
R2D2	76.5	90.2	100
D2-Net	74.8	87.0	99.8
ASLFeat	81.4	89.7	100
PosFeat	80.2	89.5	100
SuperGlue + SuperPoint	79.2	90.7	100
Sparse-NCNet	76.3	85.0	98.3
Patch2Pix	79.4	88.3	100
SNF-Feat(Ours)	81.6	90.5	100

B. Ablation Study

Building upon the feature matching experiment conducted on the Hpatches dataset, we delve deeper into understanding the contributions of non-negative learning and semantic-guided strategies within our method. For comparison purposes, we train our network using four alternative pipelines:

- Utilizing random-negative sampling strategy.
- Employing hardest negative sampling strategy, a common approach in SOTA methods.
- Integrating hardest negative sampling strategy with semantic-guided weights.
- Employing non-negative strategy without semantic-guided weights.

All other design aspects and hyperparameters remain constant across these pipelines. We evaluate the MMAScore and GPU memory usage using the same input batch size to provide a comprehensive comparison.

Results The ablation experiments presented in Table II demonstrate that our method effectively prevents model collapse without relying on negative sample pairs. When compared with the random-negative sampling method, our approach achieves superior accuracy by circumventing the disruptive effects of false-negative samples. Although the hardest-negative sampling method surpasses simple negative-free method in terms of matching ability metric, it's noteworthy that our method doesn't necessitate the extensive similarity matrix with a size of $\mathcal{O}(n^2)$ required by hardest-negative mining, thereby conserving training costs. Furthermore, our experiments validate that semantic-guided weights enhance our model's performance by suppressing the distortion in random transformed image pairs.

C. Visual Localization

Next, we evaluate SNF-Feat on the Aachen Day-Night dataset[35] for the visual localization task. This dataset provides sample codes for evaluating local features within the context of long-term visual localization. Various threshold settings are employed for evaluation, including $(0.5m, 2\circ), (1m, 5\circ), (5m, 10\circ)$.

Results. As illustrated in Table III, SNF-Feat achieves SOTA performance on the Aachen Day-Night visual localization benchmark. While some methods demonstrate comparable or slightly superior accuracy, they are often accompanied by certain limitations. For instance, SuperGlue[28] + SuperPoint[3] is a heavy and slow mixed method. ASLFeat[7] requires hardest-negative sampling and imposes high memory requirements. PosFeat[6] relies on reinforcement keypoint learning, entailing high training costs. Therefore, our method exhibits the best comprehensive performance on the visual localization task.

D. 3D Reconstruction

We assess SNF-Feat on the ETH local feature benchmark[36] for the 3D-construction task, employing

TABLE IV

3D RECONSTRUCTION RESULTS FROM DIFFERENT METHODS ON THE ETH LOCAL FEATURE BENCHMARK.

Subset	Method	Imgs	Pts	Track Length	Reproj. Err.
South Building	SuperPoint	128	159k	7.22	0.94
	RFP	128	105k	7.83	0.87
	DISK	128	120k	9.89	0.58
	PosFeat	128	139k	8.78	0.65
	SNF-Feat(Ours)	128	149k	9.69	0.58
Madrid Metropolis	SIFT + CAPS	860	245k	6.19	1.02
	SuperPoint	442	30k	9.06	1.05
	D2-Net	510	86k	6.36	1.30
	ASLFeat	620	100k	8.80	0.90
	PosFeat	407	69k	9.11	0.93
	CoAM	708	260k	6.12	1.28
	SNF-Feat(Ours)	489	95k	9.14	0.88
Gendarmenmarkt	SIFT + CAPS	1189	621k	5.35	1.01
	SuperPoint	970	95k	7.28	1.05
	D2-Net	1049	252k	5.19	1.20
	ASLFeat	1060	223k	8.62	0.94
	PosFeat	966	241k	8.45	0.99
	CoAM	1068	575k	6.69	1.29
	SNF-Feat(Ours)	979	240k	8.64	0.95
Tower of London	SIFT + CAPS	1101	450k	5.83	1.01
	SuperPoint	685	53k	8.69	0.95
	D2-Net	788	181k	5.40	1.21
	ASLFeat	823	230k	12.48	0.92
	PosFeat	779	265k	11.59	1.04
	CoAM	814	240k	5.82	1.27
	SNF-Feat(Ours)	798	256k	12.68	0.96

four metrics: the number of registered images (*Imgs.*), the number of sparse points (*Pts.*), average track length (*Track Length*), and mean re-projection error (*Reproj. Err.*).

Results. On the ETH local feature benchmark, our method achieves competitive results on the average track length and mean re-projection error metric, as outlined in Table IV. Given that our training process solely necessitates positive sample pairs generated by random homography transformations, we can leverage widely-distributed datasets. The incorporation of semantic-guided weights further enhances the authenticity of our training. Consequently, our model demonstrates proficiency in handling challenging large-scale outdoor scenes with diverse content.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we introduce a novel local feature extraction method, SNF-Feat. By proposing negative-sample-free representation learning from image-level to dense patch-level, we train discriminative feature representations while effectively preventing model collapse. Consequently, we mitigate the disturbance caused by potential false-negative samples and notably reduce memory usage. Furthermore, we incorporate an off-the-shelf semantic parsing model to enhance the authenticity of our training data while preserving generalization ability. We accomplish a self-supervised learning paradigm for the convenience of utilizing any datasets. Experimental results demonstrate that SNF-Feat outperforms SOTA methods comprehensively.

In the future, we intend to explore better keypoint localization strategies to enhance the meaningfulness of spatial components represented by our features.

ACKNOWLEDGMENT

This paper is supported by the National Natural Science Foundation of China under Grants (62073245, 62173248, 62233013). Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100) and the Fundamental Research Funds for the Central Universities. (Corresponding author: Chengju Liu, Qijun Chen)

REFERENCES

[1] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Lecture notes in computer science*, vol. 3951, pp. 404–417, 2006.

[2] R. Mur-Artal, J. M. Montiel, and J. D. Tardos, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[3] D. Detone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-Supervised Interest Point Detection and Description," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2018-June, pp. 337–349, 2018.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[5] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable CNN for joint description and detection of local features," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 8084–8093, 2019.

[6] K. Li, L. Wang, L. Liu, Q. Ran, K. Xu, and Y. Guo, "Decoupling Makes Weakly Supervised Local Feature Better," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 838–15 848.

[7] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6589–6598.

[8] C. Choy, J. Park, and V. Koltun, "Fully Convolutional Geometric Features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8958–8966.

[9] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised Learning," 2020.

[10] X. Chen and K. He, "Exploring Simple Siamese Representation Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 750–15 758.

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

[12] Y. Ono, E. Trulls, P. Fua, and K. M. Yi, "LF-Net: Learning Local Features from Images," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018.

[13] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2527–2536.

[14] X. Zhao, X. Wu, J. Miao, W. Chen, P. C. Y. Chen, and Z. Li, "ALIKE: Accurate and Lightweight Keypoint Detection and Descriptor Extraction," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.

[15] J. Revaud, P. Weinzaepfel, C. de Souza, and M. Humenberger, "R2D2: Repeatable and reliable detector and descriptor," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[16] M. J. Tyszkiewicz, P. Fua, and E. Trulls, "DISK: Learning local features with policy gradient," *Advances in Neural Information Processing Systems*, vol. 2020-Decem, no. NeurIPS, pp. 1–12, 2020.

[17] Q. Wang, X. Zhou, B. Hariharan, and N. Snavely, "Learning Feature Descriptors Using Camera Pose Supervision," in *Computer Vision – ECCV 2020*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 757–774.

[18] X. Zhou, Q. Yan, C. Liu, and Q. Chen, "PUW-Feat: A Progressive and Unified Method for Weakly Supervised Local Feature Learning," in *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2023, pp. 3795–3800.

[19] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[20] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020, pp. 1597–1607.

[21] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[22] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 9912–9924.

[23] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information theory*, vol. 37, no. 1, pp. 145–151, 1991.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[25] S. Li, Q. Yan, X. Zhou, D. Wang, C. Liu, and Q. Chen, "NDNet: Spacewise Multiscale Representation Learning via Neighbor Decoupling for Real-Time Driving Scene Parsing," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2022.

[26] 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.

[27] 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*, vol. 30. Curran Associates, Inc., 2017.

[28] P. E. Sarlin, D. Detone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning Feature Matching with Graph Neural Networks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 4937–4946, 2020.

[29] I. Rocco, R. Arandjelović, and J. Sivic, "Efficient neighbourhood consensus networks via submanifold sparse convolutions," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 2020, pp. 605–621.

[30] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8922–8931.

[31] Q. Zhou, T. Sattler, and L. Leal-Taixe, "Patch2pix: Epipolar-guided pixel-level correspondences," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4669–4678.

[32] O. Wiles, S. Ehrhardt, and A. Zisserman, "Co-attention for conditioned image matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 920–15 929.

[33] 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*, 2017, pp. 5173–5182.

[34] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[35] Z. Zhang, T. Sattler, and D. Scaramuzza, "Reference pose generation for long-term visual localization via learned features and view synthesis," *International Journal of Computer Vision*, vol. 129, pp. 821–844, 2021.

[36] 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*, 2017, pp. 1482–1491.