2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

AWENet: A Self-Supervised Network for Efficient Interest Point Detection and Description

Abstract— **We introduce AWENet (Attention-guided Wavelet Enhancement Network), an efficient self-supervised network for joint interest point detection and description that balances computational speed with feature accuracy. The network preserves fine structural details while employing multi-scale attention to enhance the discriminability of descriptors, leading to more precise and reliable interest point correspondences. Evaluations on the HPatches dataset demonstrate that AWENet achieves competitive performance in repeatability, localization accuracy, and matching robustness. Its lightweight design ensures fast processing and low computational cost, making it well-suited for applications where efficiency is critical. Qualitative results show that the network generates dense and accurate correspondences under diverse transformations, including changes in viewpoint and illumination. Overall, AWENet provides a practical and effective solution for learning local features, achieving strong matching performance without relying on heavy computation.**

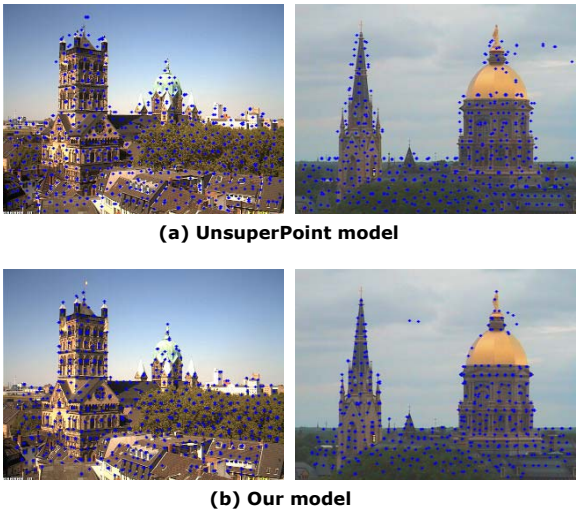
I. INTRODUCTION

Interest point detection and description aim to identify reliable 2D interest points along with representative descriptors, enabling the association of image points corresponding to the same 3D point across multiple views. Local feature detection and description are fundamental to many computer vision tasks, including visual localization [1], [2], structure-from-motion (SfM) [3], and simultaneous localization and mapping (SLAM) [4].

Traditional methods rely on hand-crafted features, focusing on low-level cues such as edges, gradients, and corners. Although these methods are designed to be robust under variations in viewpoint, scale, illumination, and noise [6], [11], [12], [14], they are limited in capturing high-level contextual information.

With the advent of deep learning, learning-based approaches have been proposed to exploit the representational power of CNNs for feature extraction. Early patch-based methods learn local descriptors from image patches around detected interest points [7], [10], [16], [17], [18], but are often hindered by independent detectors and patch-level constraints, resulting in inaccurate interest point localization and unreliable matching.

To address these limitations, recent research has shifted towards end-to-end frameworks that jointly learn both detection and description from shared CNN feature maps, particularly those extracted from deeper layers [5], [8], [13], [15]. However, existing end-to-end approaches still face challenges: high computational cost, limited descriptor discriminability under unsupervised training, and suboptimal interest point localization accuracy. As illustrated in Fig. 1, many interest points detected by UnsuperPoint [5] are slightly offset from the true salient regions, while our proposed **AWENet** produces interest points that are more accurately localized. In this work, we propose **AWENet**, a lightweight, wavelet-enhanced, multi-scale attention network with multi-objective distillation, which simultaneously improves efficiency, interest point precision, and descriptor robustness in an self-supervised learning setting.



(a) UnsuperPoint model

(b) Our model

Fig. 1. Comparison with UnsuperPoint, where blue points indicate detected interest points. No NMS was applied in either method. (a) UnsuperPoint, (b) our AWENet. While UnsuperPoint detects many interest points, some are slightly offset from true salient regions. Our method produces more accurately localized interest points, especially in complex or ambiguous areas.

II. RELATED WORK

Modern local feature learning methods have evolved from classical hand-crafted descriptors [6], [14] to patch-based learning descriptors [10], [16], [17], and further to more sophisticated feature learning approaches aimed at improving the discriminability and robustness of local descriptors. However, these methods often prioritize higher matching accuracy and robustness at the cost of significantly increased computational overhead, making them inefficient even on systems with moderate GPU resources. Moreover, they usually require extensive adaptation for different downstream tasks, limiting their deployment in large-scale applications such as visual localization [26], simultaneous localization and mapping (SLAM) [4], and structure-from-motion (SfM) [3].

In the line of joint detection and description, D2-Net [13] proposes to simultaneously perform detection and descrip-

tion within a single CNN framework, while ASLFeat [15] introduces a multi-level learning mechanism based on the D2-Net backbone to enhance low-level feature details. SuperPoint [8] employs an encoder–decoder architecture with pseudo ground-truth labels to jointly learn the detector and descriptors, eliminating the need for expensive manual annotations. Nevertheless, SuperPoint still suffers from high computational cost when applied to common image resolutions. To address this, UnsuperPoint [5] introduces a self-supervised framework trained with a Siamese scheme to automatically learn scores, locations, and descriptors of interest points from unlabeled images, while IO-Net [9] further improves robustness by incorporating an outlier rejection loss. More recently, LANet [27] adopts a pyramid strategy similar to ASLFeat to enhance descriptor representations across multiple scales, improving feature robustness without significantly increasing computational overhead. ZippyPoint [25] introduces quantization and binarization operations in CNNs to achieve significant speedup, but its reliance on custom compilation and specialized hardware instructions limits cross-platform applicability.

III. METHOD

We propose **AWENet**, as illustrated in Fig. 2(a). In the following, we provide a detailed explanation of the overall architecture and its key components.

A. Accelerated Convolutional Expansion (ACE)

Inspired by recent lightweight local feature extractors such as [21], we redesign the earliest stage of the backbone to alleviate computational bottlenecks at full image resolution. For an input image $I \in \mathbb{R}^{H\times W\times 3}$, the cost of a standard convolution is

$$\text{FLOPs} = H_i \cdot W_i \cdot C_i \cdot C_{i+1} \cdot k^2, \tag{1}$$

where the term H_iW_i dominates in the early layers. Allocating large channel depth under these conditions is inefficient and often provides limited benefit for downstream interest point extraction. Directly pruning channels across the network could reduce its robustness to challenges such as illumination and viewpoint changes, as discussed in [21].

The **Accelerated Convolutional Expansion (ACE)** retains full spatial resolution while gradually expanding the channel dimension in a *computationally efficient and progressive* manner. The module stacks several 3×3 convolutions, each followed by Batch Normalization and a lightweight learnable activation function (see Fig. 2(b)). Channels are expanded according to the following schedule:

$$3 \to 4 \to 8 \to 16 \to 32 \to 32.$$

This nonlinear expansion scheme aims to rapidly build a feature foundation with sufficient representational capacity at minimal initial computational cost, establishing the basis for accurate interest point localization at the very beginning of the network.

At the transition from 16 to 32 channels, **WCD** (see Section III-B) is applied to perform a single-level Haar wavelet decomposition, halving the spatial resolution while explicitly preserving multi-frequency components. Two additional convolutions further expand the representation, producing a compact yet information-rich feature map ready for deeper encoding.

Each layer employs a custom activation function

$$\text{AdaptiveReLU}(C) : x \mapsto \gamma\,\text{ReLU}(\alpha x) + \beta, \tag{2}$$

where α, β, γ are channel-wise learnable parameters. This formulation preserves the sparsity of ReLU while providing per-channel scaling and shifting, accelerating the adaptation of representational capacity and compensating for potential representation loss caused by aggressive computation reduction at high resolution.

By limiting channel growth before the first downsampling and inserting a frequency-preserving wavelet transform at an optimal point, **ACE** significantly reduces FLOPs in the earliest stage while retaining edge and texture cues crucial for local feature extraction. This *computationally efficient expansion* strategy ensures sufficient representation capacity for subsequent layers without incurring the heavy computational burden of early-stage high-resolution convolutions, directly contributing to speed improvement.

B. Wavelet Convolution Downsampling (WCD)

The discrete Haar wavelet transform (DWT) for image processing is extensively discussed in MWCNN [19]. While Haar DWT preserves edge and corner cues efficiently, its fixed filters cannot adapt to varying image content or task-specific patterns in self-supervised learning.

To address this, we integrate learnable convolutional features with fixed wavelet sub-bands in a unified downsampling block. The wavelet branch captures multi-frequency structural information, while the convolutional branch adaptively extracts task-relevant patterns. Together, they retain detailed spatial information and improve interest point detection in an self-supervised setting.

Given input $X \in \mathbb{R}^{C\times H\times W}$, single-level Haar DWT produces one low-frequency sub-band $y_L \in \mathbb{R}^{C\times H/2\times W/2}$ and three high-frequency sub-bands y_{HL}, y_{LH}, y_{HH}. We concatenate all sub-bands:

$$y_{\text{wave}} = \text{Concat}(y_L, y_{HL}, y_{LH}, y_{HH}) \in \mathbb{R}^{4C\times H/2\times W/2}, \tag{3}$$

then apply a 1×1 convolution with BN and ReLU:

$$y_{\text{wavelet}} = f_{\text{conv}}^{(1\times 1)}(y_{\text{wave}}) \in \mathbb{R}^{C\times H/2\times W/2}. \tag{4}$$

In parallel, a learnable stride-2 convolution produces:

$$y_{\text{stride}} = f_{\text{conv}}^{(s=2)}(X) \in \mathbb{R}^{C\times H/2\times W/2}. \tag{5}$$

Finally, the two branches are concatenated and projected via another 1×1 convolution:

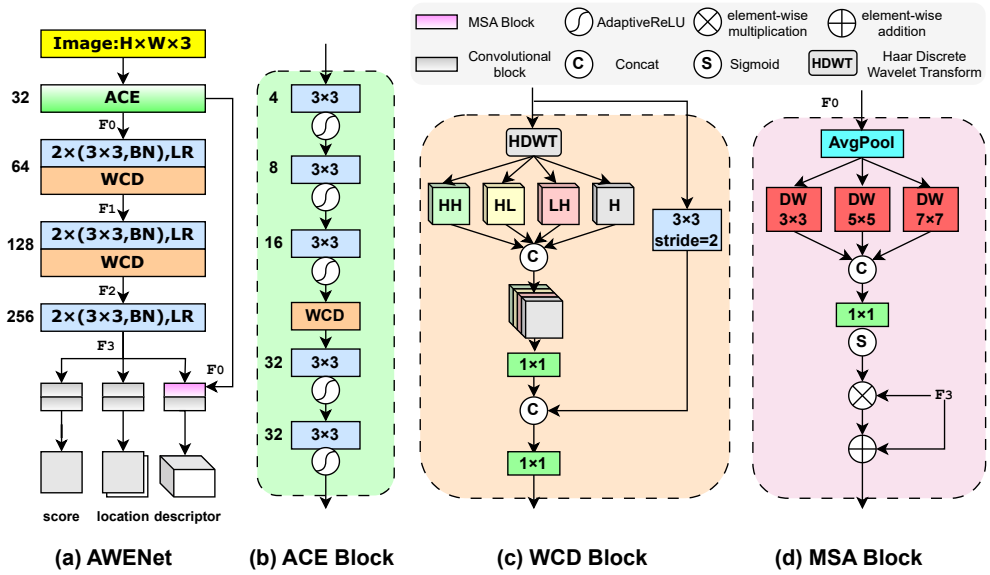


Fig. 2. Overview of the proposed AWENet. (a) Overall network architecture with an encoder–decoder design. The encoder is composed of ACE blocks and several 3×3 convolution layers, each followed by batch normalization and a LeakyReLU activation. The decoder consists of three branches: a score decoder, a position decoder, and a descriptor decoder. (b) ACE block structure. (c) WCD block for wavelet-based downsampling. (d) MSA block for multi-scale self-attention.

$$O = f_{\text{fuse}}\Big(\text{Concat}(y_{\text{wavelet}}, y_{\text{stride}})\Big) \in \mathbb{R}^{C' \times H/2 \times W/2}. \quad (6)$$

This design fuses structural cues from wavelet transform with adaptive, learnable representations, preserving fine-grained edges and corners while maintaining efficiency for subsequent interest point detection.

C. Multi-scale Attention Block (MSA)

To enrich descriptor representations while maintaining computational efficiency, we propose a *Multi-scale Attention Block*, designed to integrate the intermediate features from ACE with deeper descriptor features. A depthwise separable convolution decomposes a standard convolution into two stages: a spatial convolution applied channel-wise (depthwise), followed by a 1×1 convolution (pointwise) for channel integration. This decomposition significantly reduces FLOPs while still capturing fine-grained spatial patterns, as demonstrated in Xception [20].

In our design, the output of ACE is initially:

$$x_{\text{aceb}} \in \mathbb{R}^{32 \times H/2 \times W/2}. \quad (7)$$

To match the spatial resolution of deeper descriptor features, we first apply average pooling:

$$x_1 = \text{AvgPool}(x_{\text{aceb}}) \in \mathbb{R}^{32 \times H/8 \times W/8}. \quad (8)$$

The downsampled intermediate feature x_1 is then fed into three parallel depthwise separable convolutions with kernel sizes 3×3, 5×5, and 7×7, respectively, to capture local structural patterns at multiple receptive fields. The three outputs are concatenated along the channel dimension, followed by a 1×1 convolution, batch normalization, and Sigmoid activation, to generate the multi-scale attention map:

$$\text{attn} = \sigma\big(f_{\text{fusion}}\big(\text{Concat}(x_3, x_5, x_7)\big)\big), \quad (9)$$

where $\sigma(\cdot)$ denotes the Sigmoid activation and f_{fusion} is the 1×1 fusion convolution.

This attention map is aligned with the deeper descriptor feature

$$x_2 \in \mathbb{R}^{256 \times H/8 \times W/8}, \quad (10)$$

and applied in a residual modulation manner:

$$x_2' = x_2 + x_2 \odot \text{attn}, \quad (11)$$

where $\odot$ denotes element-wise multiplication. This operation acts as a "feature amplifier" driven by the detailed information from ACE, strengthening the spatial positions in the deep feature map that correspond to potential interest points while suppressing non-key regions. This multi-scale guidance from early-stage features effectively enhances the discriminative power of the descriptors, making their responses to interest points more accurate and prominent, while the use of depthwise separable convolutions ensures computational efficiency.

D. Self-supervised Learning with Multi-Objective Distillation

To further enhance the performance of the model under an self-supervised paradigm and to produce more discriminative and robust interest point scores, positions, and descriptors, we propose a *multi-objective knowledge distillation*

strategy [22]. This method leverages a pre-trained, high-performing SuperPoint [8] as a teacher model to guide the learning of the student model. By transferring the teacher's intermediate representations, the performance of the student model is significantly improved.

Unlike traditional distillation methods that rely on large-scale teacher networks, we employ a strong and reliable SuperPoint as the teacher to ensure efficiency and practicality during distillation. During training, the same batch of input images is fed simultaneously into the frozen teacher model and the student model. The overall loss function consists of the original self-supervised loss $\mathcal{L}_{\text{task}}$ along with multiple distillation losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{\text{score}}\mathcal{L}_{\text{score}} + \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{desc}}\mathcal{L}_{\text{desc}},$$

where $\mathcal{L}_{\text{task}}$ is the original self-supervised loss (e.g., photometric error or consistency), and $\mathcal{L}_{\text{score}}$, $\mathcal{L}_{\text{pos}}$, and $\mathcal{L}_{\text{desc}}$ are the distillation losses for score maps, position offsets, and descriptors. The coefficients λ_{score}, λ_{pos}, and λ_{desc} balance the contributions of each term.

a) Decoding and Aligning Teacher Outputs: We decode the pre-trained SuperPoint heatmap to match the output format of the student model. Specifically, the 65-channel SuperPoint score map is first normalized using softmax, and the background channel is removed. The remaining channels are then reshaped into an 8×8 grid representing local probability distributions. Within each coarse cell, the probability-weighted expected coordinates are computed to obtain sub-pixel position offsets $(\Delta x, \Delta y)$, and the maximum probability within the cell is taken as the coarse confidence score. This procedure produces a score map $(1, H/8, W/8)$ and position offsets $(2, H/8, W/8)$ that are fully aligned with the student model's outputs.

b) Score Distillation: Binary Cross-Entropy (BCE) is applied between the student and teacher score maps, weighted by high-confidence masks M. Here, the superscripts A and B indicate the two input images processed simultaneously by the model, as is standard in self-supervised interest point learning:

$$M_s^A = \mathbb{I}(S_T^A > \tau), \quad M_s^B = \mathbb{I}(S_T^B > \tau),$$

$$\begin{aligned}\mathcal{L}_{\text{score}} = &\frac{\sum \left(BCE(S_S^A, S_T^A) \odot M_s^A\right)}{\sum M_s^A + \epsilon} \\ &+ \frac{\sum \left(BCE(S_S^B, S_T^B) \odot M_s^B\right)}{\sum M_s^B + \epsilon}\end{aligned} \tag{12}$$

where $\tau = 0.05$, $\odot$ denotes element-wise multiplication, and ϵ prevents division by zero.

c) Position Distillation: L1 loss is applied to the sub-pixel position offsets, masked by teacher confidence:

$$\begin{aligned}\mathcal{L}_{\text{pos}} = &\frac{\sum \left(\|P_S^A - P_T^A\|_1 \odot M_s^A\right)}{\sum M_s^A + \epsilon} \\ &+ \frac{\sum \left(\|P_S^B - P_T^B\|_1 \odot M_s^B\right)}{\sum M_s^B + \epsilon}\end{aligned} \tag{13}$$

d) Descriptor Distillation: L2 loss (MSE) is applied per channel for descriptors:

$$\begin{aligned}\mathcal{L}_{\text{desc}} = &\frac{\sum \left(\|D_S^A - D_T^A\|_2^2 \odot M_s^A\right)}{\sum M_s^A + \epsilon} \\ &+ \frac{\sum \left(\|D_S^B - D_T^B\|_2^2 \odot M_s^B\right)}{\sum M_s^B + \epsilon}\end{aligned} \tag{14}$$

This multi-objective, mask-weighted distillation loss efficiently transfers the teacher's knowledge to the student model, significantly improving interest point detection and description performance, while fully maintaining the original self-supervised training paradigm. This strategy can be combined with any of the previously introduced network architecture improvements.

IV. EXPERIMENTS

A. Implementation Details

We implement our method using PyTorch and train it in an self-supervised manner using 118k color (RGB) images randomly sampled from the MS-COCO 2014 [23] training set. Input images are resized to 240×320 and normalized to the $[0, 1]$ range. Training uses SGD for 10 epochs (batch size 32, learning rate 0.1 with decay at 60% and 80%, weight decay 1×10^{-4}, gradient clipping at 10). The total loss consists of the original unsupervised loss $\mathcal{L}_{\text{task}}$, as defined in [5], together with the distillation losses. As specified in Section III-D, we set the distillation loss weights to $\lambda_{\text{score}} = 0.8$, $\lambda_{\text{pos}} = 1.0$, and $\lambda_{\text{desc}} = 2.5$.

B. Comparison

We evaluate the proposed AWENet on the HPatches dataset [24], which comprises 116 scenes with substantial variations in illumination and viewpoint. Our evaluation utilizes widely adopted metrics, including Repeatability (Re), Localization Error (LE), Homography Estimation Accuracy under multiple pixel thresholds ($\varepsilon = 1$, 3, and 5), and Matching Score (MS) [8]. All metrics are computed based on the top P interest points selected according to their confidence scores. Experiments are conducted at two image resolutions: 240×320 with $P = 300$, and 480×640 with $P = 1000$. The results of AWENet and several baseline methods are summarized in Table I. Additionally, we report the Mean Matching Accuracy (MMA) [13] with an error threshold of 3 pixels in Table II.

Repeatability and localization error. Repeatability measures the probability that the same interest points are detected across images, while localization error evaluates their pixel-level accuracy. As Table I shows, although the repeatability of AWENet is not the highest, it achieves the lowest localization error, demonstrating accurate and stable detection.

TABLE I

COMPARISONS ON HPATCHES DATASET WITH REPEATABILITY (RE), LOCALIZATION ERROR (LE), HOMOGRAPHY ESTIMATION ACCURACY (H-1/3/5), AND MATCHING SCORE (MS).

Methods	240×320, 300 points						480×640, 1000 points					
	Re↑	LE↓	H-1↑	H-3↑	H-5↑	MS↑	Re↑	LE↓	H-1↑	H-3↑	H-5↑	MS↑
ORB [14]	0.532	1.429	0.131	0.422	0.540	0.218	0.525	1.430	0.286	0.607	0.710	0.204
BRISK [11]	0.566	1.077	0.414	0.767	0.826	0.258	0.505	1.207	0.300	0.653	0.746	0.211
SIFT [12]	0.451	0.855	**0.622**	0.845	0.878	0.304	0.421	1.011	**0.602**	0.833	0.876	0.265
LF-Net(in) [17]	0.486	1.341	0.183	0.628	0.779	0.326	0.467	1.385	0.231	0.679	0.803	0.287
LF-Net(out) [17]	0.538	1.084	0.347	0.728	0.831	0.296	0.523	1.183	0.400	0.745	0.834	0.241
SuperPoint [8]	0.631	1.109	0.491	0.833	0.893	0.318	0.593	1.212	0.509	0.834	0.900	0.281
UnsuperPoint [5]	0.645	0.832	0.579	0.855	0.903	0.424	0.612	0.991	0.493	0.843	0.905	0.383
IO-Net [9]	**0.686**	0.970	0.591	0.867	0.912	0.544	**0.684**	0.970	0.564	0.851	0.907	0.510
Zippypoint [25]	0.682	0.813	0.503	0.862	0.916	**0.582**	0.681	0.864	0.501	**0.872**	**0.924**	**0.553**
AWENet	0.652	**0.783**	0.563	**0.873**	**0.933**	0.567	0.615	**0.862**	0.551	0.863	0.903	0.543

TABLE II

COMPARISON OF MMA@3 ACROSS DIFFERENT METHODS UNDER ILLUMINATION AND VIEWPOINT CHANGES.

	SIFT [12]	D2Net (SS) [13]	D2Net (MS) [13]	SuperPoint [8]	ASLFeat [15]	Ours
Illumination	0.525	0.568	0.468	0.738	–	**0.743**
Viewpoint	0.540	0.354	0.385	**0.639**	–	0.613
Overall	0.533	0.457	0.425	0.686	**0.723**	0.662

Homography estimation accuracy. Homography accuracy evaluates descriptor matching under geometric transformations. Using nearest-neighbor matching and RANSAC, AWENet achieves top accuracy at thresholds of 3 and 5 pixels, demonstrating geometrically consistent correspondences.

Matching score. The matching score reflects the fraction of correct correspondences among all matches. AWENet maintains competitive matching scores across resolutions, highlighting a robust balance between detection precision and descriptor distinctiveness.

Mean Matching Accuracy (MMA@3). MMA evaluates the fraction of interest points whose nearest neighbor descriptor in the other image is within a certain pixel threshold. Specifically, MMA@3 considers a match correct if the interest point distance is less than 3 pixels. As shown in Table II, our method achieves the highest MMA@3 under illumination changes and competitive performance under viewpoint variations, indicating reliable interest point detection and matching.

C. Ablation Study

In this section, we conduct an ablation study on the HPatches dataset to evaluate the individual contributions of each module in our proposed approach. All experiments are performed on an NVIDIA GTX 1050Ti GPU, which represents a relatively resource-constrained environment. The choice of such hardware aims to demonstrate that our method remains efficient and applicable even on low-end or less powerful devices. The corresponding results are presented in Table III.

Baseline. The baseline network without any of the proposed modules. It serves as a reference point to measure the effect of adding each component.

Ablation on the ACE block. Incorporating the ACE block (C1) noticeably improves FPS compared with the baseline. As we mentioned in Section III-A, this is achieved by reducing the number of channels in the early stage. Although some other metrics show slight decreases, they remain within an acceptable range, indicating that ACE provides significant computational acceleration without severely compromising detection and description performance.

Ablation on the WCD block. Integrating the WCD block (C2) results in more precise interest point localization, with a clear reduction in localization error. Although the wavelet transform introduces some additional parameters, its impact on overall performance is negligible, and the final results remain strong.

Ablation on the MSA block. Incorporating the MSA block (C3) improves descriptor discriminability and enhances the accuracy of interest point representations. Although it introduces additional computation, the performance gain in the matching score (MS) demonstrates that the module effectively strengthens feature representations without significantly affecting FPS.

Ablation on the distillation module. We incorporate the multi-objective knowledge distillation (C4) to further enhance the discriminability and robustness of interest point scores, positions, and descriptors. During training, the distillation losses are applied for only three epochs, allowing the student model to learn from the teacher while maintaining a

TABLE III
ABLATION STUDY OF AWENET ON HPATCHES.

	C1	C2	C3	C4	Re↑	LE↓	H-1↑	H-3↑	H-5↑	MS↑	FPS
Baseline	-	-	-	-	0.645	0.832	**0.579**	0.855	0.903	0.424	0.61
AWENet	✓	-	-	-	0.624	0.980	0.553	0.845	0.860	0.413	**1.26**
	✓	✓	-	-	0.626	0.842	0.522	0.866	0.886	0.433	1.15
	✓	✓	✓	-	0.636	0.824	0.539	0.862	0.912	**0.569**	1.13
	✓	✓	✓	✓	**0.652**	**0.783**	0.563	**0.873**	**0.933**	0.567	1.09

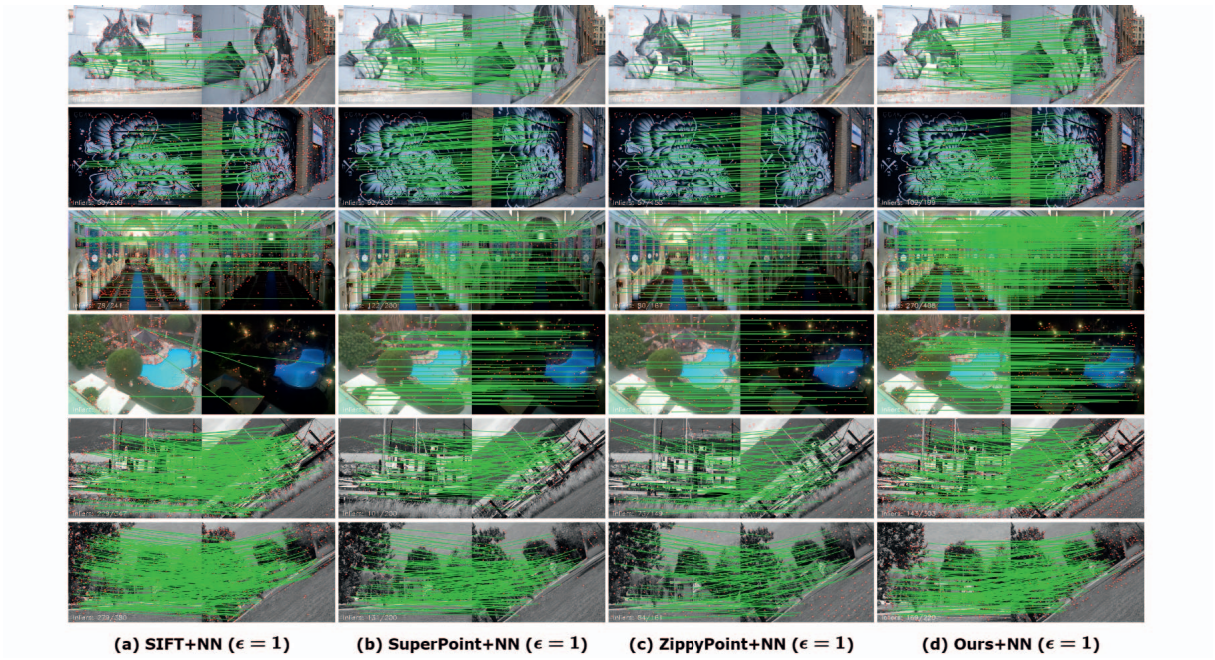


Fig. 3. Qualitative matching results on the HPatches dataset comparing SIFT+NN, SuperPoint+NN, ZippyPoint+NN, and our AWENet+NN. Top two rows: perspective transformation scenes. Middle two rows: illumination variation scenes (bright and dark). Bottom two rows: large rotation scenes. Matches are obtained using the nearest neighbor (NN) matcher with distance threshold $\epsilon = 1$.

distinction from the teacher's predictions.

D. Qualitative Matching Comparison

Fig. 3 presents qualitative matching results on the HPatches dataset, comparing SIFT+NN, SuperPoint+NN, ZippyPoint+NN, and our proposed AWENet+NN. The matches are obtained using the nearest neighbor (NN) matcher with the distance threshold $\epsilon = 1$.

The first two rows show perspective transformation scenes, where our AWENet clearly outperforms SIFT, SuperPoint, and ZippyPoint, providing more accurate and denser correspondences.

The middle two rows correspond to illumination variation scenes, including both bright and dark lighting conditions. In these scenarios, our method demonstrates strong and stable performance.

The last two rows represent images with large rotations, where SIFT exhibits the best performance due to its inherent rotation invariance.

V. CONCLUSION

We presented **AWENet**, a lightweight local feature network that balances efficiency and accuracy. Through the integration of ACE, WCD, MSA, and multi-objective distillation, AWENet achieves fast and robust feature detection and description. Experiments on HPatches demonstrate superior performance under perspective and illumination changes, making AWENet suitable for real-world applications on resource-limited devices.

REFERENCES

[1] T. Sattler et al., "Benchmarking 6dof outdoor visual localization in changing conditions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8601–8610.

[2] H. Taira et al., "Inloc: Indoor visual localization with dense matching and view synthesis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7199–7209.

[3] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[4] R. Mur-Artal, J. M. 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.

[5] P. H. Christiansen, M. F. Kragh, Y. Brodskiy, and H. Karstoft, "Unsuperpoint: End-to-end unsupervised interest point detector and descriptor," *arXiv preprint arXiv:1907.04011*, 2019.

[6] Y. Ke and R. Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, IEEE, vol. 2, 2004, pp. II–II.

[7] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 118–126.

[8] 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 workshops*, 2018, pp. 224–236.

[9] J. Tang, H. Kim, V. Guizilini, S. Pillai, and R. Ambrus, "Neural outlier rejection for self-supervised keypoint learning," *arXiv preprint arXiv:1912.10615*, 2019.

[10] P. Ebel, A. Mishchuk, K. M. Yi, P. Fua, and E. Trulls, "Beyond cartesian representations for local descriptors," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 253–262.

[11] S. Leutenegger, M. Chli, and R. Y. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *2011 International conference on computer vision*, Ieee, 2011, pp. 2548–2555.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] M. Dusmanu et al., "D2-net: A trainable cnn for joint description and detection of local features," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2019, pp. 8092–8101.

[14] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*, Ieee, 2011, pp. 2564–2571.

[15] Z. Luo et al., "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.

[16] Z. Luo et al., "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.

[17] Y. Ono, E. Trulls, P. Fua, and K. M. Yi, "Lf-net: Learning local features from images," *Advances in neural information processing systems*, vol. 31, 2018.

[18] 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.

[19] P. Liu, H. Zhang, W. Lian, and W. Zuo, "Multi-level wavelet convolutional neural networks," *IEEE Access*, vol. 7, pp. 74 973–74 985, 2019.

[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[21] G. Potje, F. Cadar, A. Araujo, R. Martins, and E. R. Nascimento, "Xfeat: Accelerated features for lightweight image matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2682–2691.

[22] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[23] T.-Y. Lin et al., "Microsoft coco: Common objects in context," in *European conference on computer vision*, Springer, 2014, pp. 740–755.

[24] 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.

[25] M. Kanakis, S. Maurer, M. Spallanzani, A. Chhatkuli, and L. Van Gool, "Zippypoint: Fast interest point detection, description, and matching through mixed precision discretization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6114–6123.

[26] 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/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 716–12 725.

[27] S. Yang, D. Zhou, J. Cao, and Y. Guo, "Rethinking low-light enhancement via transformer-gan," *IEEE Signal Processing Letters*, vol. 29, pp. 1082–1086, 2022.