Local Feature Extraction from Salient Regions by Feature Map Transformation

Yerim Jung
annie_23@inha.edu
Nur Suriza Syazwany
surizasyazwany@inha.edu
Sang-Chul Lee
sclee@inha.ac.kr

Department of Electrical
and Computer Engineering
Inha University
Incheon, South Korea

Abstract

Local feature matching is essential for many applications, such as localization and 3D reconstruction. However, it is challenging to match feature points accurately in various camera viewpoints and illumination conditions. In this paper, we propose a framework that robustly extracts and describes salient local features regardless of changing light and viewpoints. The framework suppresses illumination variations and encourages structural information to ignore the noise from light and to focus on edges. We classify the elements in the feature covariance matrix, an implicit feature map information, into two components. Our model extracts feature points from salient regions leading to reduced incorrect matches. In our experiments, the proposed method achieved higher accuracy than the state-of-the-art methods in the public dataset, such as HPatches, Aachen Day-Night, and ETH, which especially show highly variant viewpoints and illumination.

1 Introduction

Extracting and describing local features for matching is essential, especially in computer vision tasks that involve image matching, searching, tracking, and 3D reconstruction [12, 24, 39]. Feature matching focuses on three main phases when given two similar images are to be matched: feature detection, feature description, and feature matching [4, 20]. The primary goal of feature matching is to optimize matching accuracy while minimizing the memory footprint of earlier applications. The extracted features should be sparse, highly repeatable, and precise. Each image's salient features, such as its corners, are initially recognized as interest points during the detection phase. Then, local descriptors are extracted based on the neighborhood regions of these interest points and used in the matching algorithms.

Classical approaches [3, 20] concentrate on the *detect-then-describe* method, where they first detect the points by analyzing the gradient of the image before describing the points with directional information. Furthermore, these approaches [3, 20] have shifted research trends by the emergence of deep learning methods [7, 40, 44]. Since deep convolutional neural networks (DCNNs) can automatically learn features, mimic traditional detector behaviors, and process complex images, CNN methods have achieved remarkable performance than

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

before [22, 41, 42]. These data-driven methods concentrate on sparse points by leveraging descriptors' information for corresponding points. At the same time, several networks [7, 29] attempted to achieve better performance by influencing detection and description simultaneously, with improved repeatability and sparsity of detected points. Detector-free local feature matcher [38, 45] and a decoupled pipeline for a detection and description module were also studied [16].

Despite these achievements, there is insufficient consideration for light and structure information in an image. Examining this information to robustly locate and define matched points, regardless of camera viewpoint or illumination variance, is critical. Nighttime images are challenging due to the uncertainties of light and structure [38]. When viewpoints change significantly, it is also difficult to match correctly [2]. Although some studies investigated this viewpoint and light information to apply in the local feature domain, they used only hand-crafted ways such as detecting corners or simply rotating features [19, 23, 27].

In this work, we propose a new strategy that uses both style and structure information to address the issue of mismatches in image variance. Specifically, we apply the concept of Instance Selective Whitening (ISW) loss, introduced by RobustNet [6], where the features are transformed with implicit information about the style component to have robustness under variations in light. Since there are limitations in this idea and considers only the style factor, we revised ISW to consider the structure factor and apply it to the local feature field. Furthermore, we focus on salient points to reduce matching time.

Contributions. In this paper, we propose a framework that addresses the problems from a different light and structural information. We first extract features using Feature Map Generation (FMG) module. Then, Feature Map Transformation (FMT) module divides extracted features into two components: style and structure matrix. Each component independently learns the information gathered by the learned feature. Consequently, regardless of any changes in any component, the feature map will still select salient and matchable points. We introduce a loss function to maximize the influence of the structure information and minimize the style information in the feature map. The main contributions are summarized as follows:

- We overcome the limitation of feature matching in image variance by distinguishing between structure- and style-dependent features and transforming the feature maps.
- We propose Feature Map Transformation (FMT) module, exploiting an existing style transfer concept that concentrates only on style components, to transform the feature map while training to make it focus on salient features.
- Extensive experiments on different benchmark datasets demonstrate that the proposed method can achieve high accuracy in matching tasks in a short time and with fewer parameters.

2 Related Work

Local feature learning. The joint learning of feature detectors and descriptors requires a unified network to construct feature maps and allows the two tasks to share the majority of computations for improved performance. DELF [24] proposed an image retrieval technique that learns local features as a by-product of a classification loss combined with an attention mechanism to improve performance on large-scale images. It outperforms images

under changing light conditions but has limitations in terms of structural variation. SuperPoint [7] suggested a method for learning from the manual annotation of significant points on simple images such as corners and edges. However, because of their low repeatability and descriptor accuracy, it has many outliers, so the matched points tend to be mistakenly judged. R2D2 [29] has overcome this issue by learning the descriptor reliability in parallel with the detection and description phases and only selecting both repeatable and reliable keypoints with respect to the descriptor.

To find only the matchable points, D2-Net [8] proposed a describe-and-detect method for joint detection and description that uses a single CNN with shared weights. The detection is based on the entire channel's local maxima and the shape map's spatial dimensions. DISK [42] applied reinforcement learning on an end-to-end network inspired by D2-Net [8] that relied on policy gradients. Furthermore, ASLFeat [22] demonstrated significant improvement using a score map that used local shape estimation to select matching points.

Feature covariance. Previous research [9, 10] proposed that image style information is considered via feature correlations such as a gram or a covariance matrix. Since then, feature correlation has been applied to several different research areas, including style transfer [17], image-to-image translation [5], domain adaptation [30, 37], and network architecture [14, 21, 26]. Whitening transformation (WT) [5, 17, 26], which eliminates feature correlation and assigns unit variance to each feature, aids in the removal of style information from the feature representations.

Since region-specific styles and region-invariant content are simultaneously written to the covariance vector of the feature maps, whitening all the correlation components reduce feature identification and distort the boundaries of objects [17, 18]. RobustNet [6] proposed the ISW loss, which extracted only the style information to solve the problem. We want to focus on style and structure information, so we modify the ISW loss to satisfy our objective.

3 Method

3.1 Feature Map Generation Module

Feature Map Generation (FMG) module first extracts the features of an image pair, I_1 and I_2 independently, which outputs two branches: descriptors and point extraction feature maps. The point extraction branch consists of two feature maps. One produces another with a 1×1 convolution layer; the former is a reliability map $\mathbf{S}$ and the latter in repeatability map $\mathbf{R}$. The covariance matrix derived from the descriptor map $\mathbf{X}$ is used to transform the feature map to focus on saliency with style and structure information. Then FMG module uses feature maps $\mathbf{X}$, $\mathbf{S}$, and $\mathbf{R}$ to calculate the loss functions for repeatability and reliability. The FMG module's network architecture differs from R2D2 [29] but uses the same loss functions in this part. So only the architecture part will be described. The proposed method pipeline is shown in Figure 1.

As introduced in MobileNet [13], depthwise separable convolution (DSC) is a factorization convolution method that significantly reduces computation and model size with new representations. Motivated by this, we used the DSC to focus on the salient area. Inspired by [22, 29], we adopt a modified L2Net [40]—where the last layer is replaced by a set of three consecutive layers—in our backbone network to extract feature information from an image pair. Since the input images are image pairs, two backbone networks are needed. We use weight sharing in the DSC layer when adding it behind the backbone network to reduce the

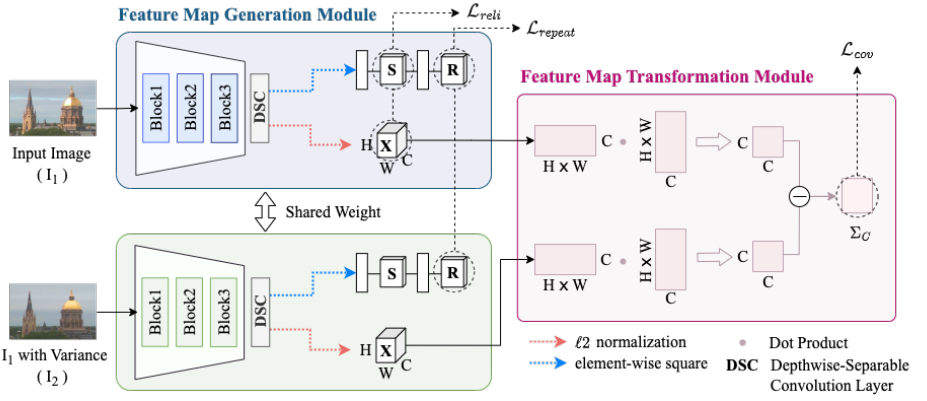

Figure 1: **Proposed framework.** Network consisting of a feature map generation module and a transformation module is shown. The reliability map **S** and repeatability map **R** learn the regions of points of interest, while the covariance matrix derived from the descriptor map **X** is used to transform the feature map to focus on saliency with style and structure information.

model weight. Furthermore, the relationship between descriptors and points of interest can be maintained by sharing the weights. The backbone network then generates three feature maps, **X** by $\ell 2$ normalization, **S** by the element-wise square, and **R** obtained from **S** with a 1×1 convolution layer. In contrast to [29], **S** and **R** come from the same branch since they depend on one another. We assumed that **S** is a feature map that learns a point that reduces the matching distance, which might affect in high repeatability rate in **R**. Therefore, the model weight gets lighter and develops more robust features through information sharing.

FMG module calculates the reliability loss, $\mathcal{L}_{reli}$, to get discriminative feature points. Let $\mathbf{X}_{ij}$ be the local descriptor in each pixel (i, j) of the image I_1; we then predict the individual reliability scores $\mathbf{S}_{ij}$ from $\mathbf{X}_{ij}$ and $\mathbf{X}^{'}_{uv}$. Here, we specify the exact coordinate (u, v) that corresponds to (i, j), knowing the ground truth correspondence mapping T, where $T \in \mathbb{R}^{H\times W\times 2}$ is the ground truth correspondence between image I_1 and I_2. $\mathbf{X}_{ij}$ is compared with $\mathbf{X}^{'}_{uv}$, where $\mathbf{X}^{'}_{uv}$ is extracted from I_2. Then, average precision is used to calculate $\mathcal{L}_{reli}$, optimized with a differentiable approximation [11, 29], using $\mathbf{S}_{ij}$.

In addition, FMG module calculates the repeatability loss, $\mathcal{L}_{repeat}$, for extracting repeatable feature points as in [29]. It uses peakiness prediction and similarity between feature pairs from input pair images. For similarity, Let **R** and $\mathbf{R}^{'}$ be the repeatability maps corresponding to I_1 and I_2. We set $\mathbf{R}^{'}_{T}$ to be a map in which $\mathbf{R}^{'}$ is transformed by the ground truth homography relationship between image pairs I_1 and I_2. Because the prime objective is to predict keypoints with high repeatability, we train the network so that the positions of the local maxima in **R** are covariant to the actual picture transformations, such as viewpoint structures or light shifts. Assuming that all the local maxima of **R** coincide with the local maxima of $\mathbf{R}^{'}_{T}$, we define the loss function. The basic concept is to maximize the cosine similarity between **R** and $\mathbf{R}^{'}_{T}$ such that the two heatmaps are identical and their maxima identically coincide. The loss may remain at a particular constant that may terminate the learning process, so we prevent this using the peakiness prediction. The final repeatability loss is calculated by considering both similarity and the peakiness of the input image pair.

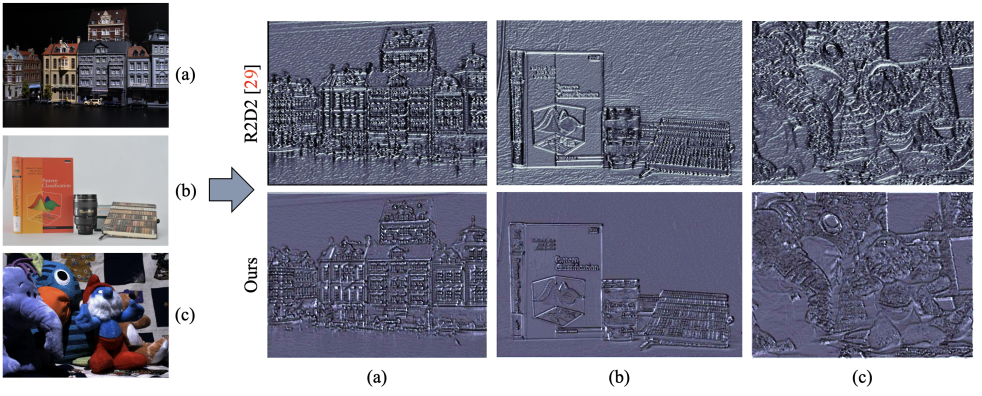

Figure 2: **Transformed feature map.** Comparison of feature maps from each network indicates the tendency in which we change the feature map; eliminates style information refers to noise, while structure complexity enlarges.

3.2 Feature Map Transformation Module

One problem with feature matching is structural perspective and lighting (style) differences. The style corresponds to noise, such as weather and light. Concentrating on the structure, which relates to the image's point of view or the edges, can result in better results. Feature Map Transformation (FMT) module suppresses the style information by performing task adaptation and supplementing the existing WT loss with an extensive transformation. We improve the previous loss by including structural information to overcome the limitations of using only existing styles.

Style/Structure Covariance Matrix. Previous studies [6, 15] claimed that applying WT to each instance of style transfer could successfully erase style information. WT is a linear transformation that equalizes the variance term in each channel to one and reduces the covariances between channels to zero. The intermediate feature map is $\mathbf{X} \in \mathbb{R}^{C \times HW}$, where C is the number of channels, and H and W are the height and width of the feature map, respectively. The covariances between pairs of channels can be defined as follows:

$$\Sigma = \frac{1}{HW}\left(\mathbf{X} - \mu \cdot \mathbf{O}^{\top}\right)\left(\mathbf{X} - \mu \cdot \mathbf{O}^{\top}\right)^{\top} \in \mathbb{R}^{C \times C} \tag{1}$$

where $\mathbf{O} \in \mathbb{R}^{HW}$ is a column vector of ones, and the μ and Σ are the mean vector and covariance matrix, respectively. When the loss function is designed so that the elements of the covariance matrix Σ decrease, the feature extraction is less affected by the style element in extracting features from the input image. This is because the feature map generation lacks the information of style elements. Based on this concept, we adopt a method of transforming feature maps using style elements. Furthermore, we use information about the structure, the leftover elements in the gram matrix beside style elements, to design the loss function in the direction of expansion rather than suppression.

Transformation Loss. In order to transform and manipulate feature maps using the aforementioned style and structure information, we introduce the FMT module which use feature map $\mathbf{X}$ to calculate $\mathcal{L}_{cov}$. This is done since the descriptor has the most information among the three output feature maps from the FMG module. FMT module then separates

style and structure characteristics from the feature representation's higher-order statistics by selectively modifying each attribute differently. Replacing the old feature map $\mathbf{X}$ with a standardized feature map $\mathbf{X}_s$ simplifies the optimization process of both the diagonal and off-diagonal elements of the covariance matrix simultaneously [6, 43]. Σ_s is a matrix made from the standardized feature map $\mathbf{X}_s$, and we define Σ_s as follows:

$$\Sigma_s = \frac{1}{HW}\mathbf{X}_s \cdot \mathbf{X}_s^\top \in \mathbb{R}^{C\times C} \tag{2}$$

After obtaining the covariance matrices $\Sigma_s(I_1)$ and $\Sigma_s(I_2)$, we calculate the difference between the two matrices to produce matrix Σ_C, as defined in Eq. 3. Matrix Σ_C indicates the sensitivity of the corresponding covariance to the photometric transformation [6]. Elements with a high variance value retain the style information, whereas elements with a low variance value retain the structure information. We use absolute value notation with vertical bars.

$$\Sigma_C = |\Sigma_s(I_1) - \Sigma_s(I_2)| \tag{3}$$

We cluster the style and structural components in equal amounts, using the mean value to determine the threshold for separating the two components. If a matrix Σ_C element is greater than the threshold, that element is classified as a style factor, while the rest of the elements are classified as structural factors. This definition is established because the prominent factors of the matrix Σ_C are assumed to imply the style factor [6]. In this case, the style factor refers to changes in light or color, and the structural factor refers to complexities with many objects, edges, or viewpoints. The proposed loss function, $\mathcal{L}_{cov}$ is formulated as follows:

$$\mathcal{L}_{cov} = \begin{cases} \mathbb{E}\left[\left\|\Sigma_C \odot \mathbf{M}_{sty}\right\|_1\right] & \text{if } \Sigma_C > \mu \\ 1 - \mathbb{E}\left[\|\Sigma_C \odot \mathbf{M}_{str}\|_1\right] & \text{otherwise} \end{cases} \tag{4}$$

where $\mathbb{E}$ and $\mathbf{M} \in \mathbb{R}^{C\times C}$ are the arithmetic mean and a mask matrix. $\mathbf{M}_{sty}$ and $\mathbf{M}_{str}$ are masks that select style and structure values, and $\odot$ is element-wise multiplication. Finally, the total loss function can be represented by Eq. 5, with each weight λ_i are empirically tuned to the optimal ratio of 1:1:2. We define $i \in \{1,2,3\}$ because the number of loss terms is three. We strengthen the argument that adding the transformation loss is superior in selecting only salient features when comparing the feature map of R2D2 [29] with ours in Figure 2.

$$\mathcal{L}_{total} = \lambda_1 \cdot \mathcal{L}_{reli} + \lambda_2 \cdot \mathcal{L}_{repeat} + \lambda_3 \cdot \mathcal{L}_{cov} \tag{5}$$

This aggregated loss function is used to select salient points to minimize the prediction of the less informative regions, such as the sky or ground. FMT is inducing the feature map transformation so that the proposed transformation loss function could extract robust features if the location where the image is taken is in the same place, regardless of the change in structure and style.

4 Experiment

4.1 Implementation details

Training. We apply Adam to optimize the network for 25 epochs with a fixed learning rate of 0.0001, a weight decay of 0.0005, and a batch size of 8 pairs of cropped images of

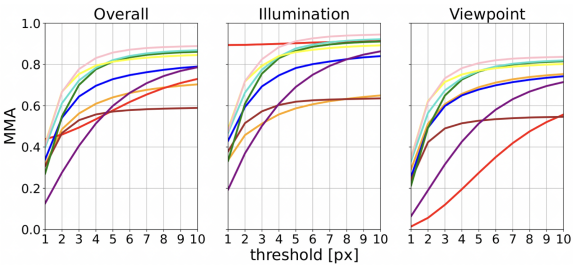

Figure 3: **Quantitative Results on HPatches.** Comparison in terms of MMA on the HPatches dataset.

	MMA@3		
	Overall	Illumi.	Viewp.
Hes.Aff. [23]	56.24	51.35	60.79
DELF(new) [24]	49.43	**89.73**	12.02
SuperPoint [7]	64.45	69.38	59.88
LF-Net [25]	53.01	57.31	49.02
D2-Net [8]	39.76	44.99	34.91
R2D2 [29]	70.06	75.56	64.96
ASLFeat [22]	72.28	75.47	68.28
DISK [42]	75.34	79.43	71.53
Ours	**78.41**	*83.22*	**73.94**

Table 1: **MMA@3 on HPatches.** Comparison at 3px threshold.

192 by 192 pixels, as in R2D2 [29]. Our experiment used the training dataset and ground-truth correspondences used in R2D2. Since our model uses the modified version of R2D2, we trained the network from scratch. Nonetheless, we fixed the patch size N used in the repeatability loss to 16 in all training parts to improve the performance of the transformation loss.

Testing. We used the sum of different scales of images to diversify the resolution of the feature maps at test time. The descriptors were interpolated at the modified locations. This multi-scale feature extraction enables the extraction of more tentative keypoints and provides improved localization.

Experiment Settings. This study used an NVIDIA GeForce RTX 3090 GPU and CUDA toolkit, version 11.2, with Python 3 and PyTorch 1.8 in the training environment.

4.2 Feature Matching

Quantitative Evaluation. We evaluated the performance of selecting meaningful points by calculating mean matching accuracy (MMA). If the distance between the converted point and the reference point exists within the threshold based on the ground truth homography metric, the converted point is classified as a correct conversion point. Figure 3 illustrates the comparisons on the H-Patches dataset [7], with MMA measured at various error thresholds. Figure 3 was drawn from the cache data provided by the D2-Net [8] repository. The comparison was performed with DELF [24], SuperPoint [7], LF-Net [25] mono and multi-scale D2-Net [8], R2D2 [29], ASLFeat [22], DISK [42], and a hand-crafted Hessian affine detector with a RootSIFT descriptor [28]. Our network denoted as "Ours" outperformed almost all state-of-the-art networks. Regarding illumination, DELF [24] outperformed our method since it identifies key points in a low-resolution feature map with a fixed grid. However, ours still exhibited the highest performance in terms of the five-pixel threshold because of the fixed grid of keypoints without spatial variation in this subgroup. The overall scores for the illumination dataset and viewpoints and their respective individual scores are presented in Table 1, with the MMA threshold set to 3.

Qualitative Evaluation. Figure 4 shows the comparison between our baseline network R2D2 [29] and the current state-of-the-art method DISK [42] with a severe change in illumination and viewpoint in the first and second row, respectively. The experiment was done with the nearest neighborhood matching with 3 points of error threshold. When we look at the yellow box, we see that we succeeded in focusing on structural information. The structurally less focused network failed to match in this part. In the red box, it can be seen that

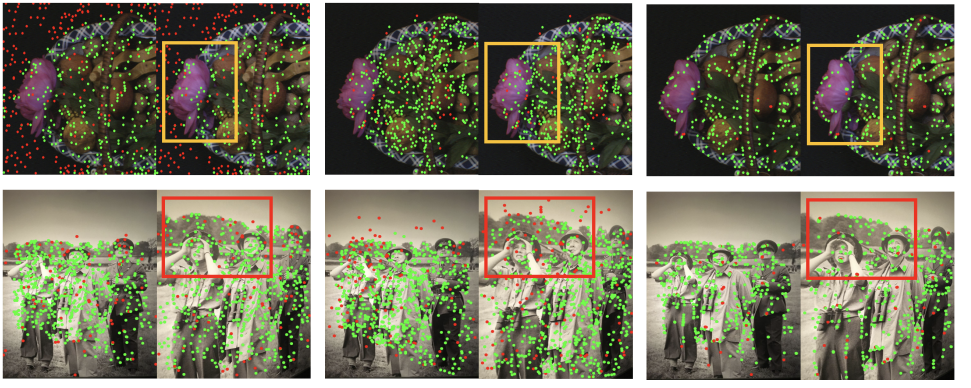

Figure 4: **Qualitative Results on HPatches.** There are illustrated pairs of illumination and viewpoint variations. Green dots represent points that are correctly matched, whereas red dots represent points that are incorrectly matched. The boxed area provides significantly better outcomes than previous methods by focusing on the salient region.

the error rate is lowered by focusing less on the background or natural objects corresponding to noise. The matching time is also shorter, shown in Figure 5.

4.3 Visual Localization

In a local reconstruction job [32], we evaluate our technique on the Aachen Day-Night dataset v1.1 [33] as in D2-Net [8]. In this section, we present the results of an optical localization task. Given the daytime photographs with known camera positions, our goal is to identify the nighttime image of the same area. The known location of the daytime photos in each set is used to triangulate the 3D structure of the scene after extensive feature matching. Finally, these 3D models are used to locate the query photographs taken at night. We followed the guidelines for Visual Localization Benchmark, for which we used our matches as input for a pre-defined visual localization pipeline based on COLMAP [34, 35]. We also adopted hierarchical localization [31] in every network for higher performance. This pipeline was then used to build an SfM model with the registered test photos. We used NetVlad [1] for the global feature and the nearest neighborhood for matching. The percentages of properly localized photos under three error levels are reported in Table 2. Our results demonstrate our method's strong generalization capabilities because of its high localization performance compared with SuperPoint [7], D2-Net [8], R2D2 [29] and DISK [42]. Our network performed fairly well when compared with matcher methods [38, 45].

4.4 3D Reconstruction

For 3D reconstruction evaluation, we used ETH-Microsoft Dataset [36] and for the evaluation protocols, we ran SfM algorithm by COLMAP [34, 35]. In Table 3, we compare our method with the results presented in [23], but only the jointly learned models. Compared with other methods, since the number of matched points was less than 100K, we do not provide the comparison with R2D2 [29]. For sparse reconstruction, we report the number of

	Day			Night		
	0.5m, 2°	1m, 5°	5m, 10°	0.5m, 2°	1m, 5°	5m, 10°
SuperPoint [8]	85.3	91.9	94.5	58.6	74.3	85.9
D2-Net [8]	81.6	89.3	96.2	62.8	80.6	92.7
R2D2 [29]	89.9	95.4	98.4	69.6	85.9	96.3
DISK [42]	-	-	-	**72.3**	86.4	**97.9**
Ours	**90.4**	**96.1**	**98.9**	**72.3**	**89.0**	*96.9*
LoFTR* [63]	-	-	-	72.8	88.5	99.0
Patch2Pix* [63]	86.4	93.0	97.5	72.3	88.5	97.9

Table 2: **Aachen Evaluation.** Comparison on Aachen for the visual localization task. * are marked for matcher methods.

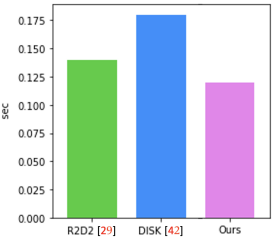

Figure 5: **Matching Time.** Comparison of matching time.

	Madrid Metropolis 1344 images				Gendarmenmarkt 1463 images				Tower of London 1576 images			
	# Reg	# Sparse	Track	Reproj	# Reg	# Sparse	Track	Reproj	# Reg	# Sparse	Track	Reproj
SuperPoint [8]	438	29K	9.03	1.02px	967	93K	7.22	**1.03px**	681	52K	8.67	0.96px
D2-Net [8]	495	**144K**	6.39	1.35px	965	310K	5.55	1.28px	708	287K	5.20	1.34px
ASLFeat [42]	649	129K	**9.56**	**0.95px**	1061	320K	**8.98**	1.05px	846	252K	**13.16**	**0.95px**
Ours	**766**	142K	8.13	1.19px	**1316**	**516K**	6.81	1.19px	**1186**	**315K**	8.63	1.21px

Table 3: **ETH Evaluation.** 3D reconstruction held with ETH-Microsoft Dataset.

registered images (# Reg), the number of sparse points (# Sparse), tracked length (Track), and reprojection error (Reproj). Table 3 data shows that our network performs favorably against previous methods on the 3D reconstruction task. Furthermore, the number of registered images obtained the best value, and the number of sparse points also had the best or second-order value. It can be interpreted that the selection of salient points was made well.

4.5 Ablation Studies

We validated the significance of the components that comprise our suggested transformation loss function by performing two ablation studies. We studied the presence of style and structural component losses to determine how each characteristic contributes to learning. $\mathcal{L}_{sty}$ and $\mathcal{L}_{str}$ denote loss function that use same subscript in the masks of Eq. 4. The result in Table 4 reveals that the matching accuracy is not only influenced by the style factor but also by the structure factor. This experiment confirms the efficiency of ISW when applied to our method and its ability to provide additional information about the structure. Furthermore, ablation studies were conducted with and without DSC layer. The performance improved using DSC layer, with a model weight reduction of 2 MB.

5 Conclusion

We proposed a robust network using self-transformation loss, which transforms a feature map that contributes to the repeatability of local features. We separated structure and style characteristics by clustering the covariance vector and influenced the feature of each characteristic. The feature maps were unified to make two feature maps closer by reducing the light component and sharpening the structure component. Consequently, this step increases the repeatability of the points, reduces outliers, and assigns robustly matched descriptors. A comparison with similar research reveals that our method more effectively extracts robust matching points in various scenes. Nevertheless, the current work has limitations. Points

	HPatches			Aachen (Day)			Aachen (Night)		
	Overall	Illum	Viewp	0.5m, 2°	1m, 5°	5m, 10°	0.5m, 2°	1m, 5°	5m, 10°
w/o $\mathcal{L}_{sty}$&$\mathcal{L}_{str}$	70.06	75.56	64.96	89.9	95.4	98.4	69.6	85.9	96.3
w/o $\mathcal{L}_{sty}$	72.08	78.04	66.55	89.8	**96.1**	98.7	**73.3**	88.5	95.3
w/o $\mathcal{L}_{str}$	76.53	81.43	71.99	89.7	95.8	98.5	69.6	84.8	96.3
w/o DSC	76.89	81.95	72.19	89.9	95.3	98.2	69.1	85.9	93.7
w/ All	**78.41**	**83.22**	**73.94**	**90.4**	**96.1**	**98.9**	*72.3*	**89.0**	**96.9**

Table 4: **Ablation Studies.** Ablation experiment on MMA@3 and visual localization to see how each component affect transformation loss. We show that the best results occur when style and structure loss are included with depth-wise convolution.

tend to be retrieved by clustering them in a particular region. This analysis reveals that adaptively selecting from sparse points is a promising avenue for future research.

6 Acknowledgment

This work was supported in part by the National Research Foundation of Korea (NRF) under Grant NRF-2021R1A2C2010893 and in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-00155915, Artificial Intelligence Convergence Innovation Human Resources Development (Inha University).

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016.

[2] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5173–5182, 2017.

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[4] Jian Cheng, Cong Leng, Jiaxiang Wu, Hainan Cui, and Hanqing Lu. Fast and accurate image matching with cascade hashing for 3d reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2014.

[5] Wonwoong Cho, Sungha Choi, David Keetae Park, Inkyu Shin, and Jaegul Choo. Image-to-image translation via group-wise deep whitening-and-coloring transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10639–10647, 2019.

[6] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via

instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11580–11590, 2021.

[7] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018.

[8] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features. *arXiv preprint arXiv:1905.03561*, 2019.

[9] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28: 262–270, 2015.

[10] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[11] Kun He, Yan Lu, and Stan Sclaroff. Local descriptors optimized for average precision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 596–605, 2018.

[12] Jared Heinly, Johannes L Schonberger, Enrique Dunn, and Jan-Michael Frahm. Reconstructing the world* in six days*(as captured by the yahoo 100 million image dataset). In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3287–3295, 2015.

[13] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[14] Lei Huang, Dawei Yang, Bo Lang, and Jia Deng. Decorrelated batch normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 791–800, 2018.

[15] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1446–1455, 2019.

[16] Kunhong Li, Longguang Wang, Li Liu, Qing Ran, Kai Xu, and Yulan Guo. Decoupling makes weakly supervised local feature better. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15838–15848, 2022.

[17] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *arXiv preprint arXiv:1705.08086*, 2017.

[18] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 453–468, 2018.

[19] Yuan Liu, Zehong Shen, Zhixuan Lin, Sida Peng, Hujun Bao, and Xiaowei Zhou. Gift: Learning transformation-invariant dense visual descriptors via group cnns. *Advances in Neural Information Processing Systems*, 32, 2019.

[20] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[21] Ping Luo. Learning deep architectures via generalized whitened neural networks. In *International Conference on Machine Learning*, pages 2238–2246. PMLR, 2017.

[22] Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Aslfeat: Learning local features of accurate shape and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6589–6598, 2020.

[23] Iaroslav Melekhov, Gabriel J Brostow, Juho Kannala, and Daniyar Turmukhambetov. Image stylization for robust features. *arXiv preprint arXiv:2008.06959*, 2020.

[24] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017.

[25] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. Lf-net: Learning local features from images. *arXiv preprint arXiv:1805.09662*, 2018.

[26] Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, and Ping Luo. Switchable whitening for deep representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1863–1871, 2019.

[27] Rémi Pautrat, Viktor Larsson, Martin R Oswald, and Marc Pollefeys. Online invariance selection for local feature descriptors. In *European Conference on Computer Vision*, pages 707–724. Springer, 2020.

[28] Michal Perd'och, Ondrej Chum, and Jiri Matas. Efficient representation of local geometry for large scale object retrieval. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 9–16. IEEE, 2009.

[29] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32:12405–12415, 2019.

[30] Subhankar Roy, Aliaksandr Siarohin, Enver Sangineto, Samuel Rota Bulo, Nicu Sebe, and Elisa Ricci. Unsupervised domain adaptation using feature-whitening and consensus loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9471–9480, 2019.

[31] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019.

[32] Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, and Tomas Pajdla. Are large-scale 3d models really necessary for accurate visual localization? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1646, 2017.

[33] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8601–8610, 2018.

[34] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[35] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016.

[36] Johannes L Schonberger, Hans Hardmeier, Torsten Sattler, and Marc Pollefeys. Comparative evaluation of hand-crafted and learned local features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1482–1491, 2017.

[37] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[38] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021.

[39] Linus Svärm, Olof Enqvist, Fredrik Kahl, and Magnus Oskarsson. City-scale localization for cameras with known vertical direction. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1455–1461, 2016.

[40] Yurun Tian, Bin Fan, and Fuchao Wu. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 661–669, 2017.

[41] Yurun Tian, Vassileios Balntas, Tony Ng, Axel Barroso-Laguna, Yiannis Demiris, and Krystian Mikolajczyk. D2d: Keypoint extraction with describe to detect approach. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[42] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020.

[43] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[44] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2015.

[45] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe. Patch2pix: Epipolar-guided pixel-level correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4669–4678, 2021.