3DFA: Aligning the Features Between Point Cloud and Query Image for Scene-Specific Visual Localization

Sizhe Song[1], Yankuan Chi[2], Shuhan Zhong[1], S.-H. Gary Chan[1]
The Hong Kong University of Science and Technology
{ssongad,szhongaj,gchan}@cse.ust.hk[1], ychiad@connect.ust.hk[2]

Abstract

We consider scene-specific visual localization, i.e., estimating the camera poses of a query image taken in a specific site given the training data composed of support images in the same site and their corresponding reconstructed 3D point cloud. A critical performance factor in the localization pipeline is aligning the features of the 2D query image with the training data for the subsequent keypoint matching with the point cloud. Prior art has been primarily focusing on aligning the query images with the support images without making the most of the rich geometry and appearance information of the point cloud. To leverage the point cloud information so as to achieve higher localization accuracy, we propose 3DFA, a novel approach which directly aligns the 2D query image with the features of the 3D model. 3DFA employs a simple yet effective hierarchical patching operation designed for direct 2D-3D alignment, and a patch-matching loss without the need for any additional annotation. Our extensive experiments on representative datasets demonstrate that 3DFA outperforms the SOTA methods substantially, cutting the translation and rotation errors by 18.3% and 8.4%, respectively, and achieving robustness against different levels of scene noise.

1. Introduction

Visual localization is to estimate the camera poses of a query image taken in a 3D scene. It has important commercial applications in robotics, augmented reality, and navigation. We consider the common case of scene-specific visual localization, i.e., the neural network for localization is trained and tested at the same site or "scene" (e.g., a warehouse, mall, airport, station, etc.). As compared with its scene-agnostic counterpart based on foundation models [18, 32–34], scene-specific localization generally achieves higher accuracy due to better optimization and customization for the site (as confirmed in our extensive experiments).

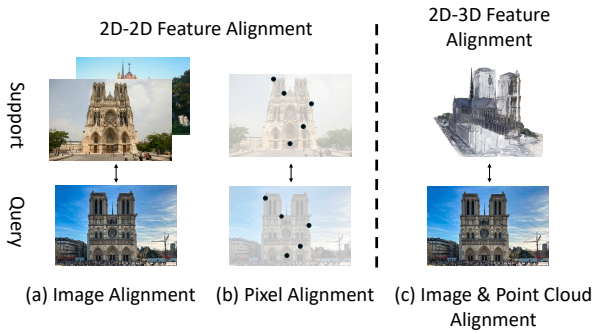

Figure 1. We propose 3DFA, a novel 2D-3D feature alignment approach to directly align the features of the 2D features of the query image with that of the 3D point cloud for scene-specific visual localization. By involving the point cloud into the feature alignment process *(c)*, 3DFA beats previous methods with only 2D-2D feature alignment at either image level or pixel level *(a, b)*. To facilitate such alignment, 3DFA employs a novel hierarchical patching operation as well as a patch-matching loss function.

The training data of a scene-specific visual localization network is composed of support images (i.e., images of known poses taken at different locations of the target scene at different times) and a point cloud built from these images with each of its point having a feature optimized in the 3D reconstruction process. After training, the current localization pipeline locates a query image in three major steps: 1) *Feature extraction:* Extract keypoints and their corresponding features from the query image as query features (using a 2D encoder); 2) *Feature alignment:* Align the query features with the features of support images (extracted using the same encoder) as the aligned keypoints of the query images; and 3) *Keypoint matching:* Match the features of the aligned keypoints of the query image with the points of the point cloud. The matched keypoint-point pairs are then input to the PnP algorithm to obtain the camera poses [13]. It is clear from the above that, to achieve accurate visual localization, feature of 2D keypoints must be similar to those of corresponding points in the 3D point cloud; this naturally requires considering the geometry and appearance informa-

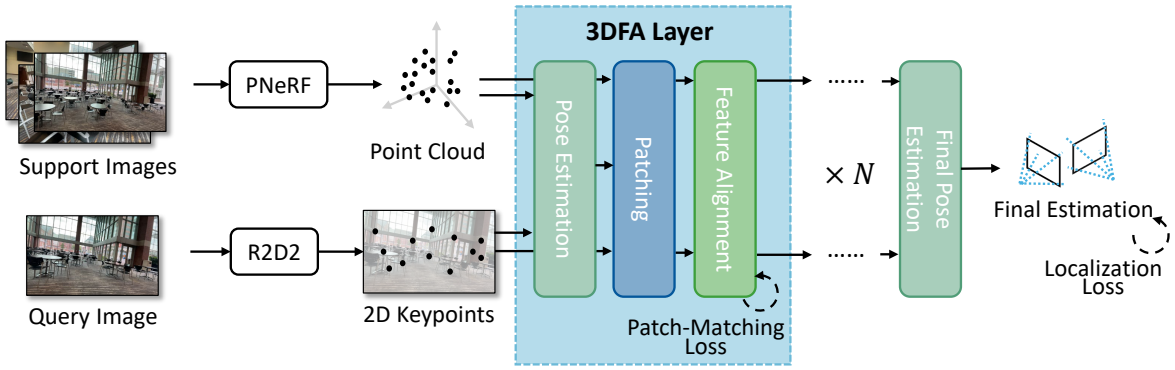

Figure 2. There are multiple 3DFA layers in the network, each consisting of 1) pose estimation; 2) *Hierarchical Patching* (Section 3.2); and 3) *Feature Alignment Module* (Section 3.2). Patch-matching loss (Section 3.3) is also included in each 3DFA layer as supervision. At the end, a final pose estimation (Section 3.4) will be conducted using aligned features.

tion of the 3D model.

Step 2 in the above pipeline, i.e. *feature alignment*, has been widely known to play a critical determinant role in the accuracy and robustness of visual localization. Without a properly designed feature alignment, the keypoint features of the query image and point cloud features would not be matched well, and this inevitably degrades localization performance. Early works on visual localization consider no feature alignment (i.e., no Step 2) but simply match the extracted keypoint features with the reconstructed point cloud features [4, 5, 29, 30, 39, 40, 42, 44]. Because the features of the query images and point cloud are extracted and optimized in different ways, the matched pairs are not satisfactory, hence degrading the performance. More recent ones study 2D-2D feature alignment, where they align the image or pixel features between the query and support images [20, 25]. Because 3D features have not been considered in their alignment, these methods have not fully utilized the information in the point cloud, compromising the keypoint matching and localization performance.

To overcome the shortcoming of 2D-2D feature alignment, we propose, for the first time, a scene-specific 2D-***3D F***eature ***A***lignment network that directly aligns the features of the 2D query image with that of 3D point cloud for visual localization. The network, termed ***3DFA***, is briefly compared with previous approaches in Figure 1: we replace the support images generally used in previous approaches by the point cloud to leverage its rich geometry and appearance information. Designing such 2D-3D feature alignment presents us two key challenges. First, how to effectively relate the 2D features with the 3D ones? Second, how to design a proper loss function to optimize alignment without any extra annotation effort?

To respond to these challenges, 3DFA makes the following two contributions:

- **A hierarchical patching operation to align 2D and 3D features:** We propose a hierarchical patching operation that divides the 2D query image and 3D point cloud into spatially relevant patches. After establishing the corresponding 2D and 3D patch pairs, we apply feature alignment directly on each pair.
- **A novel patch-matching loss function without additional annotation effort:** To optimize feature alignment, we propose a patch-matching loss function. The function utilizes the correspondence established by hierarchical patching without any extra annotation cost.

We conduct extensive experiments on representative datasets that cover diverse scenes and scales. Our experiments demonstrate that 3DFA achieves significant performance improvement over the state of the art, reducing the median localization errors by 18.3% and 8.4% in translation and rotation, respectively. 3DFA is also robust against different levels of noise in the scene.

2. Related Work

2.1. Visual Localization

Early works directly use 2D images to regress camera poses [1, 6, 12, 15, 16, 24, 27, 31, 35, 36, 41, 43], since these methods do not have explicit 3D representations in their pipelines, they are widely outperformed by the more common ones with a 3D model. Given support images of the target scene, visual localization first reconstructs a 3D model with features, usually a point cloud, and then extracts keypoints with features from the query image. Finally, it matches 2D keypoints to 3D points and uses PnP algorithm to output the camera pose [8, 11, 13]. Some of these works do not employ any feature alignment in their networks, instead they explore different combinations of 3D reconstruction methods as well as 2D keypoint extraction ways [2–5, 7, 10, 14, 17, 21, 22, 26, 29, 30, 39, 40, 42, 44]. Later, methods with 2D-2D feature alignment are proposed

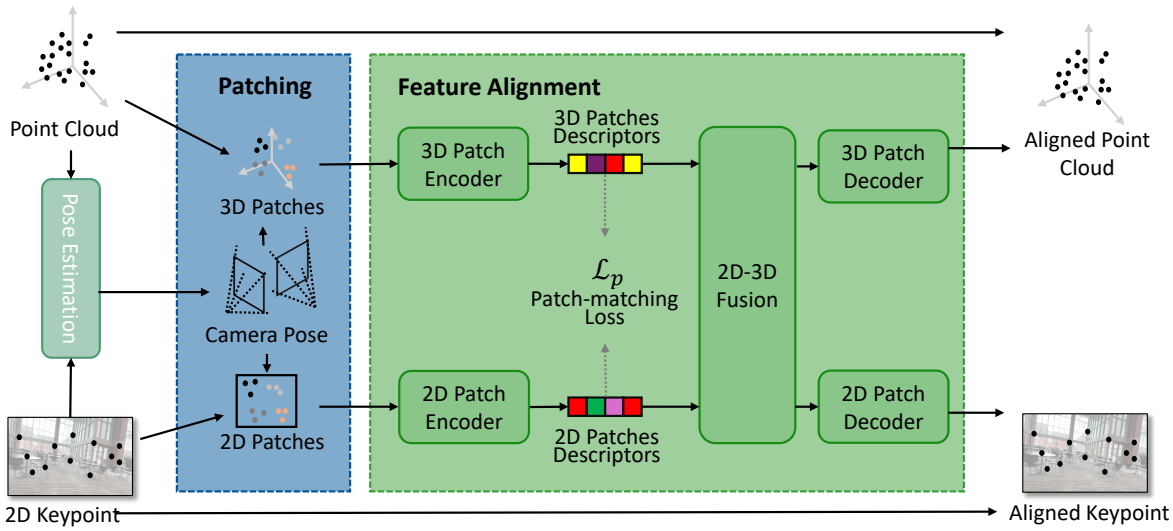

Figure 3. In one 3DFA layer, we utilize the estimated camera pose to divide 2D and 3D features into spatially corresponding patches. Then feature alignment is applied to each pair of patches and outputs the updated 2D and 3D features.

to improve localization performance by aligning query image features to the support images [20, 25]. While these methods address image-level misalignments arising from variations in illumination and style, they neglect to incorporate point cloud features during feature alignment, therefore failing to leverage the rich geometry and appearance information embedded within the point cloud. Lately, a series of foundation models have also been presented to jointly tackle multiple 3D tasks through pure regression [18, 32–34]. These models are often trained with a massive amount of high-quality data and hence have good generalizability. While they are also able to output camera poses as one of the general tasks, their accuracy is commonly lower than dedicated scene-specific visual localization approaches (as confirmed in our experiments).

2.2. 2D-2D Feature Alignment

Existing works bypass a direct 2D-3D feature alignment and resort to a 2D-2D one. Their feature alignment is conducted on image or pixel level between the query image and the support images. One work focuses on the potential image style variation. It aligns query image feature to support image features through image style transfer [20]. The other work tries to align individual pixels between query image and support images [25]. These works achieve better performance compared with similar networks without alignment; however, they do not directly exploit the geometry and appearance information in the point cloud. Hence, their performance is still suboptimal.

3. 3DFA Scheme

3.1. Pipeline Overview

We are given a support image set $\mathcal{I}_s$ of a 3D scene, their camera poses that consist of rotation matrices and translation vectors $R_s, t_s \in SE(3)$ and the camera intrinsic parameters K. Our goal is to compute the poses of a query image $\mathcal{I}_q$ such that the differences between estimated pose $\hat{R}_q, \hat{t_q}$ and ground truth pose R_q, t_q are minimized.

Figure 2 illustrates the pipeline of our proposed 3DFA network. We begin with 3D scene reconstruction using the support image set $\mathcal{I}_s$ and ground truth camera parameters R_s, t_s, K. We choose PNeRF as our 3D representation [37], a point-cloud-based approach with good performance. Meanwhile, PNeRF is a very general point cloud, each point is a 3D coordinate plus a single feature vector. Hence, 3DFA will also enjoy the advancement of more sophisticated future point-based representations, for example any 3D Gaussian splatting variants. The 3D reconstruction step will eventually return us a point cloud $\mathcal{P} = \{p_1, p_2, \cdots\}$ where each individual point $p_i = \{c_i \in \mathbb{R}^3, f_i \in \mathbb{R}^d\}$ contains a 3D coordinate plus a d-dimensional feature vector.

The next step is to extract keypoints from the query image $\mathcal{I}_q$. For this task, we employ the widely adopted R2D2 network [23], which is known for its robust performance in generating reliable keypoints. This step does not rely on any specific characteristics of the R2D2 network so that 3DFA is, again, compatible with any alternatives. The result of this step is a keypoint set $\mathcal{K} = \{k_1, k_2, \cdots\}$ where each individual keypoint $k_i = \{C_i \in \mathbb{R}^2, F_i \in \mathbb{R}^d\}$ contains a 2D coordinate plus a feature vector.

Then we move to our proposed 2D-3D feature alignment. The point cloud $\mathcal{P}$ and 2D keypoints $\mathcal{K}$ will together go through multiple rounds of feature alignment. Each round contains the pose estimation, the patching operation and the alignment. After this hierarchical and iterative feature alignment, the final aligned 2D and 3D features are used to estimate the output $\hat{R}_q, \hat{t}_q$.

During training, support images with ground truth camera poses are used as the query to train 3DFA network. In the inference stage, unseen query images are used to evaluate network performance.

3.2. 2D-3D Feature Alignment

Hierarchical Patching. We need correspondence between features in order to align them. The intuition of patching is that spatially corresponding regions in the image and point cloud are the ones we should align. Such feature alignment ensures keypoints are more likely to match with the point cloud correctly, thereby improving localization accuracy. This is possible by utilizing the camera parameters. Specifically, we can project a 3D point at coordinate c to the image plane and derive its image coordinate C as

$$\begin{bmatrix} C \\ 1 \end{bmatrix} = K \cdot [R \mid t] \cdot \begin{bmatrix} c \\ 1 \end{bmatrix}, \tag{1}$$

where R, t, K are camera poses and intrinsic parameters. Note that in our pipeline, pose estimation is done before patching, which aims to produce a camera pose for the patching operation. During training, we can ignore the estimated pose and directly use ground truth pose to enhance learning.

This projection operation establishes a mapping relationship between each 3D point and its corresponding 2D region. To preserve image locality and facilitate efficient computation, we then divide the image and point cloud into corresponding patches based on the projection. Specifically, we divide the query image into square patches, then identify which keypoints are inside a patch and which 3D points are projected to that patch. Note that our design is also compatible with other patching shapes and areas, if necessary.

Depending on the specific patch partition used, one single patching operation only focuses on a specific scale of the features. In order to extend our approach to multiple scales, we employ a hierarchical structure, which includes multiple patching operations in total, each working with a different patch size. Hierarchical patching helps to facilitate feature alignment at different scales, iteratively improve feature quality and enhance pose estimation.

Feature Alignment Module. After one patching operation, we have multiple pairs of 2D and 3D patches. Denote a specific pair, say the i^{th} one, as $\langle \mathcal{P}_i, \mathcal{K}_i \rangle$, where $\mathcal{P}_i \subseteq \mathcal{P}$ and $\mathcal{K}_i \subseteq \mathcal{K}$ are two subsets of 3D points and 2D keypoints. We start by encoding the 2D and 3D patch descriptors. For simplicity of notation, we omit the index i for all descriptors in this subsection. Then $h_{\mathcal{P}}$ and $h_{\mathcal{K}}$ are

$$h_{\mathcal{K}} = \phi_1 \left(\frac{\sum_{k_j \in \mathcal{K}_i} F_j}{\|\mathcal{K}_i\|} \right), \tag{2}$$

$$h_{\mathcal{P}} = \phi_2 \left(\frac{\sum_{p_j \in \mathcal{P}_i} f_j}{\|\mathcal{P}_i\|} \right), \tag{3}$$

where ϕ_1, ϕ_2 are the 2D and 3D patch encoder respectively. The patch encoders are implemented by MLPs so the output descriptors are also feature vectors. The extracted patch descriptors aggregate and encapsulate the features of all points within a patch. Subsequently, we align the 2D and 3D patch descriptors by

$$h'_{\mathcal{K}} \oplus h'_{\mathcal{P}} = \phi_3 \Big(h_{\mathcal{K}} \oplus h_{\mathcal{P}} \Big), \tag{4}$$

where $h'_{\mathcal{K}}, h'_{\mathcal{P}}$ are the updated patch descriptors after feature alignment, ϕ_3 are the 2D-3D fusion layer and $\oplus$ represents concatenation. This step enables feature exchange between the corresponding 2D and 3D patch descriptors, ensuring that both descriptors hold semantically similar features to their spatially associated regions. Finally, the updated patch descriptors are decoded and integrated back into the original 2D and 3D features to complete the feature alignment

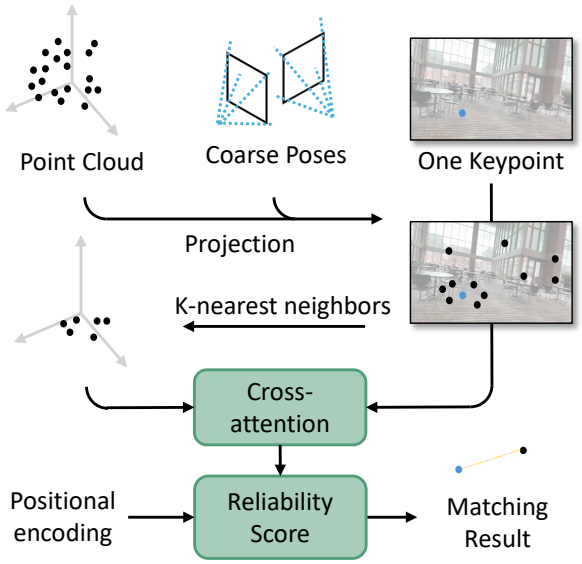

Figure 4. The final pose estimation module. It first determines the final keypoint matching result in a small 3D neighborhood by a cross-attention layer, it then removes those low-confidence matches by a reliability score. This module enables 3DFA to adjust the final output for one last time.

Table 1. Evaluation on large scenes selected from DL3DV-10K dataset. We report the median translation error / rotation error (centimeters / degrees). The SOTA performance is highlighted in bold while the second-best and a tie is marked with an underline.

Method	Mall	Hall	Store	Market	Court	Feast	Average
PoseNet	82.4/9.18	126.7/11.31	109.5/11.23	112.1/10.93	214.7/15.73	128.4/12.97	129.0/11.89
DFNet	36.7 / 1.99	62.0 / 2.68	55.3 / 2.41	56.2 / 2.42	97.6 / 3.44	128.4 / 12.97	61.1 / 2.63
DSAC*	15.4 / 1.84	27.3 / 2.43	22.4 / 2.22	22.9 / 2.31	42.5 / 3.16	27.8 / 2.40	26.4 / 2.39
PNeRFLoc	14.7 / 1.53	26.0 / 2.10	21.2 / 1.85	21.0 / 1.91	43.0 / 2.83	25.0 / 2.20	25.1 / 2.07
DUSt3R	19.9 / 2.07	34.7 / 2.81	29.1 / 2.44	28.4 / 2.61	55.2 / 3.54	33.5 / 2.99	33.4 / 2.74
MASt3R	18.4 / 1.91	31.7 / 2.54	26.7 / 2.32	26.5 / 2.31	51.4 / 3.47	31.2 / 2.70	31.0 / 2.54
VGGT	18.5 / 1.96	31.5 / 2.63	25.9 / 2.27	26.5 / 2.37	53.5 / 3.50	30.6 / 2.71	31.1 / 2.57
ACE0	15.4 / 1.60	27.2 / 2.18	22.2 / 1.93	21.9 / 1.99	44.2 / 2.95	26.1 / 2.31	26.5 / 2.16
ACE	14.6 / 1.55	25.7 / 2.11	21.0 / 1.87	20.7 / 1.93	41.9 / 2.85	24.7 / 2.24	24.8 / 2.09
PixLoc	14.0 / 1.54	24.9 / 2.05	20.6 / 1.85	21.4 / 1.92	40.5 / 2.81	25.3 / 2.19	24.5 / 2.06
NeRFLoc	13.5 / 1.53	23.9 / 2.02	19.5 / 1.84	22.3 / 1.90	39.4 / 2.77	25.6 / 2.13	24.0 / 2.03
3DFA (ours)	**11.5 / 1.42**	**19.5 / 1.85**	**17.8** / 1.84	**20.2** / 1.90	**35.4 / 2.56**	**24.4** / 2.13	**21.5 / 1.95**

process

$$F_j \leftarrow F_j + \phi_4\Big(F_j + h'_{\mathcal{K}}\Big), \tag{5}$$

$$f_j \leftarrow f_j + \phi_5\Big(f_j + h'_{\mathcal{P}}\Big), \tag{6}$$

where ϕ_4, ϕ_5 are patch decoders. After this step, the updated 2D and 3D features are ready for another round of patching and alignment, forming the hierarchical structure of 3DFA.

3.3. Patch-Matching Loss

Given the established matching relationships between 2D and 3D patches, we naturally employ a patch-matching loss $\mathcal{L}_p$ to supervise the feature alignment process. In one feature alignment module, say there are n pairs of patches in total, for the i^{th} 2D patch $\mathcal{K}_i$ and the j^{th} 3D patch $\mathcal{P}_j$, A_{ij} is the cosine similarity between their patch descriptors $h_{\mathcal{K}_i}$ and $h_{\mathcal{P}_j}$, then

$$\mathcal{L}_p = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(A_{ij} - I_{ij}\right)^2, \tag{7}$$

where I is the $n \times n$ identity matrix. This patch-matching loss is essentially forcing the differentiation of patches, which constrains the network to generate similar descriptors for corresponding patches, further enhancing feature alignment. Finally the network will be supervised together by the patch-matching loss and the traditional localization loss $\mathcal{L}_l$ [9],

$$\mathcal{L} = \alpha\mathcal{L}_p + \beta\mathcal{L}_l, \tag{8}$$

where α, β are hyperparameters.

3.4. Final Pose Estimation

We employ a fine-matching module at the last pose estimation part to further refine 3DFA's final output. As illustrated in Figure 4, for every keypoint, we find M points from the point cloud with the closest projections on the image plane, which is a very small subset in the neighborhood since the estimated pose at this moment is considered to be relatively accurate. Within this subset, we optimize keypoint matching for one last time by performing a cross-attention operation between the feature of this keypoint and the points in the neighborhood. The 3D point with the highest attention score is identified as the final matching result. And to exclude low-confidence matches, we compute a reliability score for each matched pair. Specifically, for a matched pair $\langle k_i, p_j \rangle$, the reliability score s is

$$s = \phi_6\Big(\text{PE}(k_i, p_j) \oplus (F_i, f_j)\Big), \tag{9}$$

where $\text{PE}(\cdot)$ is the commonly used 2D and 3D positional encoding function, and ϕ_6 is the reliability prediction layer.

4. Illustrative Experimental Results

4.1. Data Preparation

To thoroughly assess the effectiveness of our proposed 3DFA approach, we select three datasets featuring diverse 3D scenes of various scales: ***7Scenes*** [28] and ***ScanNet++*** [38] have room-level or below-room-level scenes. Among the extensive collection of scenes, we eventually select ten of them to represent smaller environments with a variety of styles and objects. ***DL3DV-10K*** [19] is a dataset

Table 2. Median translation error (centimeters) on ScanNet++ (left) and 7Scenes (right). SOTA performance is highlighted in bold while the second-best or a tie is marked with an underline.

Method	Work	Studio	Bedroom	Living	Business	Dorm	Heads	Office	Chess	Fire	Average
PoseNet	28.3	30.8	31.7	28.9	30.2	32.0	29	48	32	47	33.79
DFNet	13.9	15.3	15.9	14.0	15.2	15.9	3	7	4	4	10.82
DSAC*	6.0	6.4	6.7	5.9	6.4	6.8	1	3	2	2	4.62
PNeRFLoc	5.8	6.1	6.5	5.7	5.9	6.6	1	3	2	2	4.46
DUSt3R	6.5	6.9	7.5	6.9	6.8	7.6	2	3	2	3	5.22
MASt3R	6.0	6.5	7.1	6.3	6.6	7.0	1	3	2	3	4.85
VGGT	6.2	6.3	7.0	6.0	6.6	7.2	1	3	2	2	4.73
ACE0	5.9	6.2	6.7	5.7	5.9	6.4	1	3	2	2	4.48
ACE	5.9	5.8	6.5	5.6	5.8	6.2	1	3	2	2	4.38
PixLoc	5.8	6.2	6.4	6.0	6.2	6.2	1	3	2	2	4.48
NeRFLoc	5.4	5.7	6.2	5.2	5.4	6.2	1	**2**	2	2	4.11
3DFA (ours)	**4.8**	**5.2**	**5.8**	**4.7**	**4.7**	**5.6**	1	3	2	2	**3.88**

with large-scale scenes whose width is usually beyond 20 meters. Again, we select six representative scenes with diverse styles and categories. For convenience purposes, we assign a casual name to each scene from ScanNet++ and DL3DV-10K, as they are originally named by hashes.

4.2. Comparison With State-Of-The-Art

Table 1 compares the performance 3DFA scheme against other networks on DL3DV-10K. We follow the common evaluation practices to report the median translation and rotation errors of different scenes. We can see that 3DFA achieves state-of-the-art translation performance across all six scenes. Additionally, it has SOTA rotation performance in the majority of scenes, with only two of them resulting in a tie. In the end, our method also obtains the best average localization results.

Such results demonstrate the efficacy of our direct 2D-3D feature alignment and hierarchical patching operation, particularly in large-scale scenes where the point cloud is larger and keypoint matching is harder without proper feature alignment. 3DFA directly aligns features of the keypoints and point cloud based on their spatial correspondence to improve keypoint matching. Our hierarchical patching operation also facilitates such alignment across different scales, while some existing methods, instead, focus either on individual points or the entire image.

Table 2 and Table 3 list the experimental results on room-level scenes. In 7Scenes, all performances of various methods are comparable, which is not surprising in these below-room-level scenes. We can safely draw the conclusion that visual localization is solved in these small scenes. Then in Scannet++, 3DFA achieves SOTA translation results across all six scenes and SOTA rotation performance in five of them. Our strong performance in these typical real-life room level scenes highlights the potential of 3DFA in real-world applications.

During inference, 3DFA relies on the initial pose estimation to start feature alignment. We here demonstrate how robust 3DFA is against noises in the initial poses in Table 4, where the query images in *Hall* scene is split into three subsets according to difficulty (i.e. initial pose estimation error). 3DFA outperforms the other two alignment schemes in all three subsets, proving our strong robustness against inaccurate initial pose. We believe such results are contributed by our coarse-to-fine hierarchy.

4.3. Ablation Study

We conduct the ablation study on ScanNet++ by using different settings of alignment and loss, as shown by Table 5. Specifically, we remove all the 3DFA layers to obtain the first row. We keep one 3DFA layer with fixed 3×3 patch partitions to obtain the second row. We remove the patch-matching loss to obtain the third row. Finally, the last row is our complete network. By comparing the first and second rows, we observe that direct 2D-3D feature alignment improves localization performance across all four scenes. By comparing the second and the last row, we observe that further enhancement is achieved through the hierarchical structure. Note that in the *Work* and *Studio* scenes, hierarchical patching works significantly better. We attribute this result to the repeated elements in such kind of environments, where raw features of individual keypoint lead to

Table 3. Median rotation error (degrees) on ScanNet++ (left) and 7Scenes (right). The SOTA performance is highlighted in bold while the second-best or a tie is marked with an underline.

Method	Work	Studio	Bedroom	Living	Business	Dorm	Heads	Office	Chess	Fire	Average
PoseNet	5.04	4.77	6.05	4.92	5.22	5.56	12	7.7	8.1	14.4	7.38
DFNet	1.21	1.21	1.43	1.17	1.31	1.39	1.8	2	1.5	2.2	1.52
DSAC*	1.05	1.03	1.25	1.06	1.14	2.30	1.8	1.2	1.1	1.2	1.31
PNeRFLoc	0.93	0.92	1.13	0.91	0.97	1.08	0.8	1	0.8	0.9	0.94
DUSt3R	1.15	1.15	1.36	1.11	1.14	1.26	1.9	1.3	1.1	1.3	1.28
MASt3R	1.12	1.14	1.31	1.09	1.17	1.24	1.8	1.2	1.0	1.2	1.23
VGGT	1.12	1.15	1.29	1.07	1.24	1.35	1.9	1.2	1.1	1.2	1.26
ACE0	0.99	0.98	1.22	0.95	1.02	1.11	1.1	1.2	1.0	1.1	1.07
ACE	0.96	0.93	1.16	0.92	1.00	1.06	0.9	1.1	1.0	1.1	1.01
PixLoc	0.96	0.96	1.17	0.96	1.05	1.06	0.8	**0.8**	0.8	**0.7**	0.93
NeRFLoc	0.91	0.92	1.04	0.90	0.95	1.08	1.9	1.1	1.1	1.1	1.10
3DFA (ours)	**0.87**	**0.87**	**0.97**	0.90	**0.91**	**0.99**	0.8	1.1	0.8	0.9	**0.91**

Table 4. We split the scene *Hall* into three subsets by their initial estimation error and compare the final median translation and rotation error (centimeters / degrees).

	Hard	Medium	Easy
PNeRFLoc	39.2 / 2.95	26.0 / 2.10	14.2 / 0.95
PixLoc	36.4 / 2.87	24.9 / 2.05	14.0 / 0.95
3DFA (ours)	**32.7 / 2.75**	**20.1 / 1.89**	**11.3 / 0.87**

Table 5. Median translation error (centimeters) on the ScanNet++ dataset of 3DFA with different alignment and loss settings.

3DFA	Work	Studio	Bedroom	Living
w/o align	6.0	6.5	6.7	6.2
w/ one align	5.7	5.9	6.3	5.2
w/o $\mathcal{L}_p$	5.4	5.5	6.0	4.8
3DFA	**4.8**	**5.2**	**5.8**	**4.7**

ambiguities in matching. Hierarchical patching effectively addresses this issue by emphasizing feature locality. By removing the patch-matching loss, we observe a degradation in all scenes. This loss term serves as an inductive bias for our 3DFA layers, forcing the similarity between spatially corresponding patches. Our results prove its ability to enhance feature alignment.

Table 6. Median translation error (centimeters) on the ScanNet++ dataset when different patch partitions are used. For example, "3,5,7,9" means four 3DFA layers where the first one has 3×3 patches and the last one has 9×9 patches.

# of Patches	Work	Studio	Bedroom	Living
3, 3, 3, 3	6.0	6.2	6.5	6.0
3, 5, 7, 9	**4.8**	**5.2**	**5.8**	**4.7**
9, 7, 5, 3	5.0	5.3	6.1	5.0

4.4. Hyperparameter Analysis

4.4.1. Patch Partition

We investigate different combinations of patch partitions in our 3DFA layers. Table 6 presents the results on four ScanNet++ scenes. We test a uniform, an ascending and a descending style. It is observed that the ascending number of patches yields the best localization performance. We attribute this to the fact that unaligned 3D features typically exhibit smaller receptive fields and focus more on local details. By starting with broader contextual information and progressively moving to fine-grained details, 3DFA becomes more effective. On the other hand, a descending number of patches, while suboptimal, still improves localization accuracy compared to using a fixed number of patches. This is because our hierarchical patching works at different scales, extracting and aligning comprehensive keypoint features. Limiting this process to a single scale neglects valuable information at other scales. The uniform

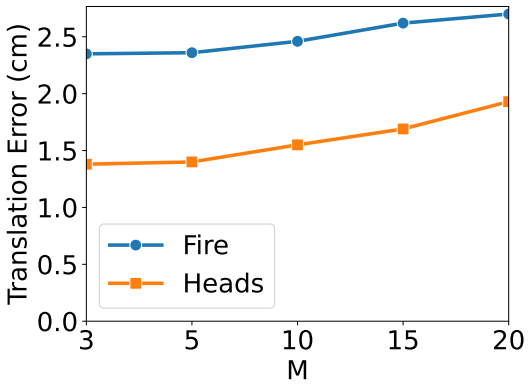

Figure 5. Translation error (centimeters) with different final pose estimation scope M.

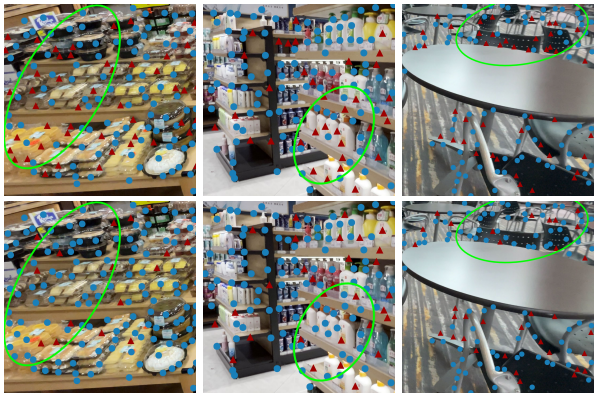

Figure 6. The initial keypoint match result (left) vs the final result after all 3DFA layers (right). The red triangles are bad matches and blue circles are good ones. The ellipses mark the areas that are significantly improved by our feature alignment. The three examples are extracted from the DL3DV-10K dataset.

patch partition leads to a bad performance which is even worse than second row of Table 5. We believe such repetitive partitioning leads to the overfiting to a certain scale and the absence of all other scales, which eventually degrades the localization ability.

4.4.2. Final Pose Estimation Scope M

We analyze different scope, M, for the KNN algorithm in the final pose estimation part. As illustrated in Figure 5, in the *Fire* and *Heads* scenes, the optimal value of M lies between 3 and 5, and larger values gradually degrade matching performance. Such result is not surprising, since 3DFA has iteratively aligned 2D and 3D features for multiple rounds, the estimation is very close to ground truth such that only a minor refinement within a very small scope is necessary. A larger M in this step will include more points, which on the contrary introduces noises back to the aligned features.

4.5. Case Study

Figure 6 illustrates several representative cases of initial and final match results in 3DFA. Note that our 2D-3D feature alignment provides crucial geometric constraints that help differentiate between visually similar keypoints, such as those in the ellipses. These ambiguous keypoints are a major source of localization error of previous methods that rely primarily on appearance features. In contrast, 3DFA incorporates point cloud geometry, and through multiple rounds of iterative feature alignment, the model progressively refines these matches by leveraging spatial coherence, resulting in substantially improved matching accuracy as demonstrated in the final outputs.

5. Conclusion

We propose, for the first time, to directly align the features between 3D point cloud and 2D query image for scene-specific visual localization. Our scheme, termed 3DFA, employs a hierarchical patching operation and patch-matching loss to effectively take into account the rich geometry and appearance features of the 3D model reconstructed from support images. 3DFA substantially boosts the performance of visual localization as compared with the SOTA 2D-2D alignment approaches, reducing median errors of translation and rotation by 18.3% and 8.4%, respectively. Our work demonstrates the critical role of explicit 2D-3D alignment for accurate scene-specific visual localization. Given its good localization performance, 3DFA holds significant promise for deployment in application domains like indoor robotic navigation and augmented reality, where agents operate reliably within the kinds of complex, pre-mapped environments, represented exactly by our evaluation datasets that covers small to large scales.

Acknowledgement This work was supported, in part, by Research Grants Council Collaborative Research Fund (under grant number C1045-23G) and RGC-General Research Fund (under grant number 16201625) of Hong Kong.

References

[1] Vassileios Balntas, Shuda Li, and Victor Prisacariu. Relocnet: Continuous metric learning relocalisation using neural nets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 751–767, 2018. 2

[2] Eric Brachmann and Carsten Rother. Learning less is more-6d camera localization via 3d surface regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4654–4662, 2018. 2

[3] Eric Brachmann and Carsten Rother. Visual camera relocalization from rgb and rgb-d images using dsac. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5847–5865, 2021.

[4] Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. Accelerated coordinate encoding: Learning to relocalize in minutes using rgb and poses. In *CVPR*, 2023. 2

[5] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Aron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. In *European Conference on Computer Vision*, pages 421–440. Springer, 2024. 2

[6] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2616–2625, 2018. 2

[7] Federico Camposeco, Andrea Cohen, Marc Pollefeys, and Torsten Sattler. Hybrid scene compression for visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7653–7662, 2019. 2

[8] Changhao Chen, Bing Wang, Chris Xiaoxuan Lu, Niki Trigoni, and Andrew Markham. Deep learning for visual localization and mapping: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[9] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2781–2790, 2022. 5

[10] Shuai Chen, Xinghui Li, Zirui Wang, and Victor A Prisacariu. Dfnet: Enhance absolute pose regression with direct feature matching. In *European Conference on Computer Vision*, pages 1–17. Springer, 2022. 2

[11] Andy Couturier and Moulay A Akhloufi. A review on deep learning for uav absolute visual localization. *Drones*, 8(11): 622, 2024. 2

[12] Mingyu Ding, Zhe Wang, Jiankai Sun, Jianping Shi, and Ping Luo. Camnet: Coarse-to-fine retrieval for camera relocalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2871–2880, 2019. 2

[13] MA FISCHLER. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981. 1, 2

[14] Khang Truong Giang, Soohwan Song, and Sungho Jo. Learning to produce semi-dense correspondences for visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19468–19478, 2024. 2

[15] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5974–5983, 2017. 2

[16] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 2

[17] Minjung Kim, Junseo Koo, and Gunhee Kim. Ep2p-loc: End-to-end 3d point to 2d pixel localization for large-scale visual localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21527–21537, 2023. 2

[18] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 1, 3

[19] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 5

[20] Jianlin Liu, Qiang Nie, Yong Liu, and Chengjie Wang. Nerf-loc: Visual localization with conditional neural radiance field. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9385–9392. IEEE, 2023. 2, 3

[21] Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. Coordinet: uncertainty-aware pose regressor for reliable vehicle localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2229–2238, 2022. 2

[22] Vojtech Panek, Zuzana Kukelova, and Torsten Sattler. Meshloc: Mesh-based visual localization. In *European Conference on Computer Vision*, pages 589–609. Springer, 2022. 2

[23] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32, 2019. 3

[24] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2

[25] Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, et al. Back to the feature: Learning robust camera localization from pixels to pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3247–3257, 2021. 2, 3

[26] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1744–1756, 2016. 2

[27] Yoli Shavit, Ron Ferens, and Yosi Keller. Learning multi-scene absolute pose regression with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2733–2742, 2021. 2

[28] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013. 5

[29] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6825–6834, 2022. 2

[30] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7199–7209, 2018. 2

[31] Mehmet Ozgur Turkoglu, Eric Brachmann, Konrad Schindler, Gabriel J Brostow, and Aron Monszpart. Visual camera re-localization using graph neural networks and relative pose supervision. In *2021 International Conference on 3D Vision (3DV)*, pages 145–155. IEEE, 2021. 2

[32] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21686–21697, 2024. 1, 3

[33] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[34] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 1, 3

[35] Bowen Wen and Kostas Bekris. Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8067–8074. IEEE, 2021. 2

[36] Dominik Winkelbauer, Maximilian Denninger, and Rudolph Triebel. Learning to localize in new environments from synthetic training data. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5840–5846. IEEE, 2021. 2

[37] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5438–5448, 2022. 3

[38] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 5

[39] Hongjia Zhai, Xiyu Zhang, Boming Zhao, Hai Li, Yijia He, Zhaopeng Cui, Hujun Bao, and Guofeng Zhang. Splatloc: 3d gaussian splatting-based visual localization for augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 2

[40] Hongjia Zhai, Boming Zhao, Hai Li, Xiaokun Pan, Yijia He, Zhaopeng Cui, Hujun Bao, and Guofeng Zhang. Neuraloc: Visual localization in neural implicit map with dual complementary features. *arXiv preprint arXiv:2503.06117*, 2025. 2

[41] Wei Zhang and Jana Kosecka. Image based localization in urban environments. In *Third international symposium on 3D data processing, visualization, and transmission (3DPVT'06)*, pages 33–40. IEEE, 2006. 2

[42] Boming Zhao, Luwei Yang, Mao Mao, Hujun Bao, and Zhaopeng Cui. Pnerfloc: Visual localization with point-based neural radiance fields. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7450–7459, 2024. 2

[43] Qunjie Zhou, Torsten Sattler, Marc Pollefeys, and Laura Leal-Taixe. To learn or not to learn: Visual localization from essential matrices. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3319–3326. IEEE, 2020. 2

[44] Qunjie Zhou, Maxim Maximov, Or Litany, and Laura Leal-Taixé. The nerfect match: Exploring nerf features for visual localization. In *European Conference on Computer Vision*, pages 108–127. Springer, 2024. 2