BoQ: A Place is Worth a Bag of Learnable Queries

Supplementary Material

6. More implementation details

Training image size. For our training implementation, we resized the images to 320×320 pixels to maintain consistency with the training procedures of [2–4]. This choice of specific resolution directly influences the spatial resolution of the resulting feature maps, which is 20×20 when using a cropped ResNet backbone [24]. This choice also allows for the use of larger batch sizes.

Inference image size. Considering that many benchmarks contain images of varying sizes and aspect ratios—and often at higher resolutions— we resize the images to a resolution slightly higher than 320 pixels while preserving the original aspect ratio. This approach maintains the integrity of the scenes by keeping the original aspect ratio, and allows the learned queries to interact with bigger feature maps. In Tab. 9 we perform testing at different image sizes (288, 320, 384, 432 and 480), using a BoQ model trained with 320×320 images. As we can see, when the images are resized to a height of 384 pixels, there is a consistent improvement in Recall@1 across almost all datasets. This experiment suggests that heights of 384 and 432 may represent an optimal balance between image detail and the model's capacity to extract and utilize informative features. Note that the performance gains from resizing to these heights are marginal compared to the baseline size of 320 pixels.

Inference im. height	AmsterTime	Eynsham	St-Lucia	SVOX (all)	Nordland**
288	48.3	90.6	99.8	98.5	74.8
320	50.0	91.0	99.7	98.5	75.4
384	53.1	91.3	99.9	98.5	73.4
432	51.4	91.5	99.5	98.6	70.0
480	50.1	91.6	99.4	98.5	65.9

Table 9. Recall@1 performance on various benchmarks, testing with images resized to varying heights (in pixels) while preserving their original aspect ratio. The model was trained on GSV-Cities using a fixed image size of 320×320.

Data Augmentation. For data augmentation, we adopted the same strategy used in [3, 4], employing RandAugment [16] with $N = 3$, which specifies the number of random transformations to apply sequentially.

Training Time. The training of our model, incorporating a ResNet-50 backbone [24] and two BoQ blocks (as depicted in Fig. 4) on GSV-Cities dataset [3], takes approximately 6 hours on a 2018 NVidia RTX 8000, with the power consumption capped at 180 watts.

7. Vision Transformer backbones

For this experiment, we trained our technique using two distinct Vision Transformer backbones, each coupled with a single BoQ block, and trained on GSV-Cities dataset [3]. The first backbone, ViT [19], pretrained base variant with 86M parameters, of which we froze all but the last two blocks to allow for fine-tuning. This model is designed to process fixed input image sizes of 224×224. The second backbone, DinoV2 [37], also with 86M parameters, underwent a similar process of freezing, leaving the final two blocks unfrozen for training. DinoV2+BoQ was trained with images resized to 280×280 and tested with images resized to 322×322.

The results presented in Tab. 10 indicate the impact of Vision Transformer backbone on BoQ model's performance. For ViT+BoQ, performance was possibly hindered by ViT's fixed input size of 224×224 leading to performance degradation. This is particularly noticeable on MSLS, SPED and AmsterTime datasets, where the Recall@1 performance is notably lower than ResNet-50 + BoQ. In contrast, DinoV2+BoQ pushes the boundaries to achieve new state-of-the-art results. The increased Recall@1 scores across all benchmarks are substantial, especially on MSLS, Pittsbug, SPED and AmsterTime. These substantial gains underscore DinoV2's capability in image feature extraction, thereby enhancing BoQ model's performance.

	MSLS	Pitts250k	Pitts30k	SPED	AmsterTime	Eynsham	SVOX
ViT+BoQ	87.6	93.9	91.0	83.9	44.3	88.7	97.6
DinoV2+BoQ	92.9	95.8	93.3	91.9	60.2	95.5	98.6

Table 10. Recall@1 performance of BoQ coupled with Vision Transformer backbones. ViT uses a fixed input size of 224×224 which may explain the performance decline. In contrast, with DinoV2 as backbone, we achieve new state-of-the-art scores on every benchmark, with significant margins.

8. Interpretability of the learned queries

In this section, we demonstrate how the learned queries in BoQ can be visually interpreted through their direct interactions with the feature maps via cross-attention mechanisms. To do so, we examine their behavior in images of the same location under viewpoint changes, occlusions, and varying weather conditions.

The cross-attention mechanisms in our BoQ model have been instrumental in achieving fine-grained feature discrimination, as demonstrated by Fig. 5, Fig. 6 and Fig. 7. These

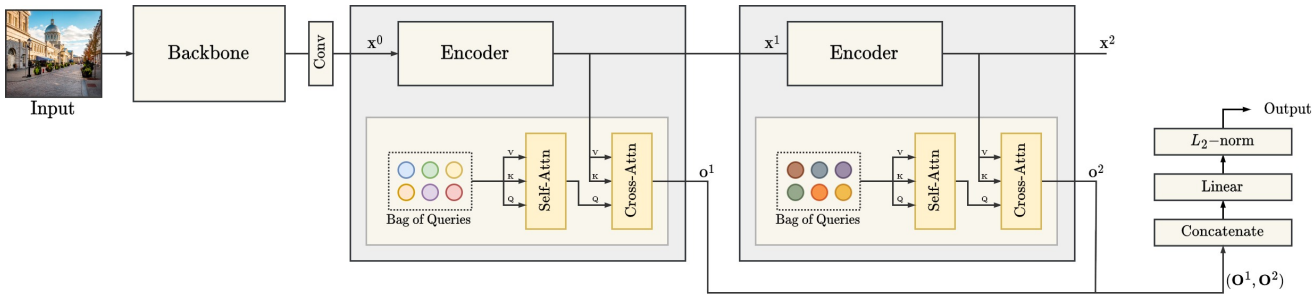

Figure 4. Detailed architecture of our model using ResNet-50 backbone and two BoQ blocks.

figures provide a visualization of the learned queries' attention patterns across diverse urban scenes and under various environmental conditions.

Fig. 5 demonstrates the model's temporal robustness, displaying consistent attention across images of the same location captured at different times. The learned queries reliably focus on distinctive features, such as buildings, foliage, and poles, despite variations in viewpoint, lighting, and weather conditions.

Moving objects within a scene often pose a challenge for VPR techniques. Nonetheless, as shown in Fig. 6 our learned queries focus their attention towards static elements of the environment, avoiding moving objects like vehicles.

Fig. 7 underscores the specialization of the learned queries, showcasing their selective focus on relevant features, such as vegetation and buildings. This selective attention is indicative of our model's ability to interpret complex visual information within the environment.

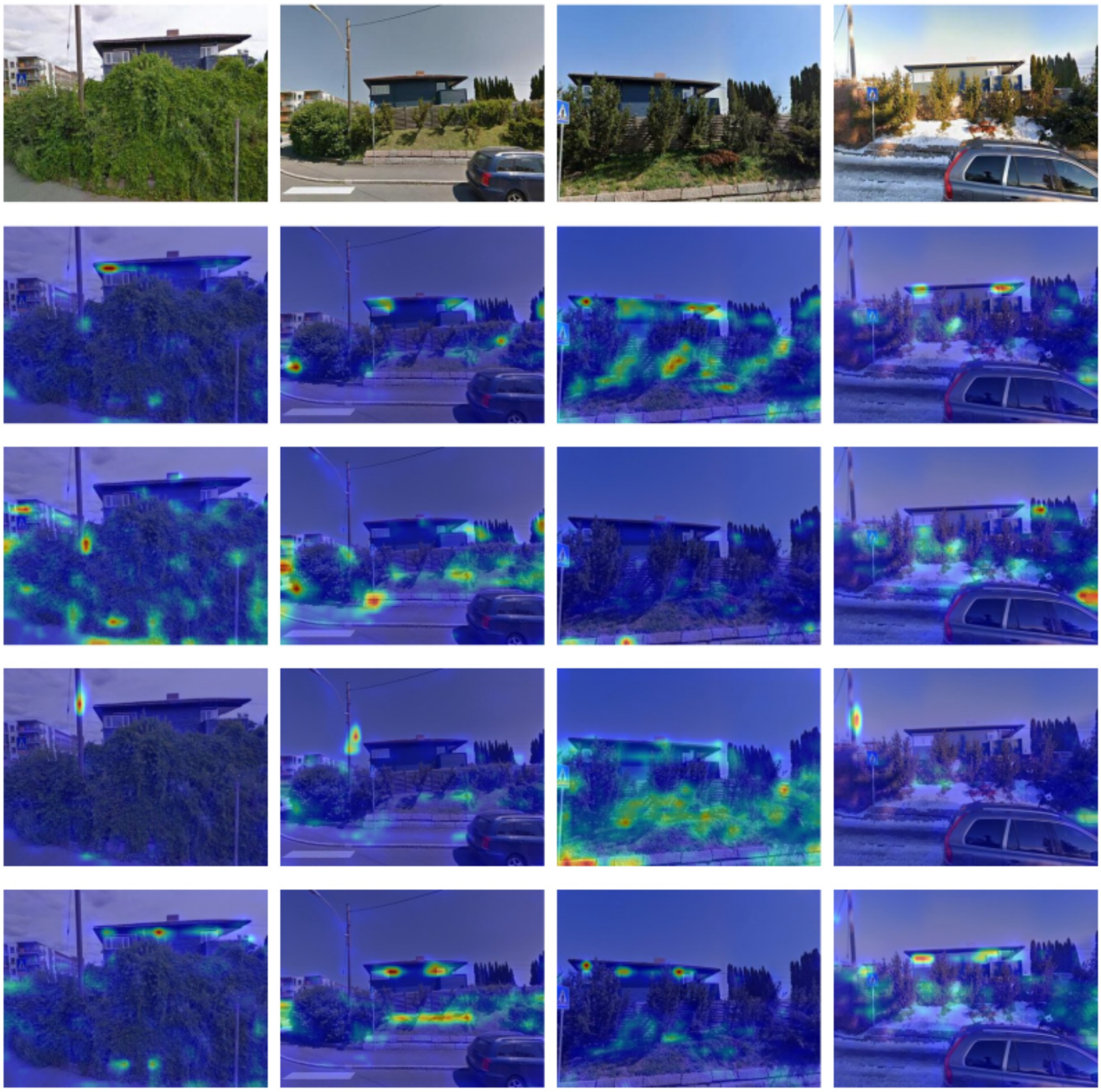

Figure 5. **Weather and occlusions.** The first row displays four images of the same location captured at different times, illustrating changes in the environment. Subsequent rows reveal the cross-attention scores between one learned query and the feature maps of the respective input image. In these heatmaps, regions with higher attention scores are indicated in warmer colors (red/yellow), signifying areas where the query is focusing more intensely. First row shows four images of the same place accross different times. The following four rows show the cross-attention scores of four selected learned queries on the feature maps of the input image.

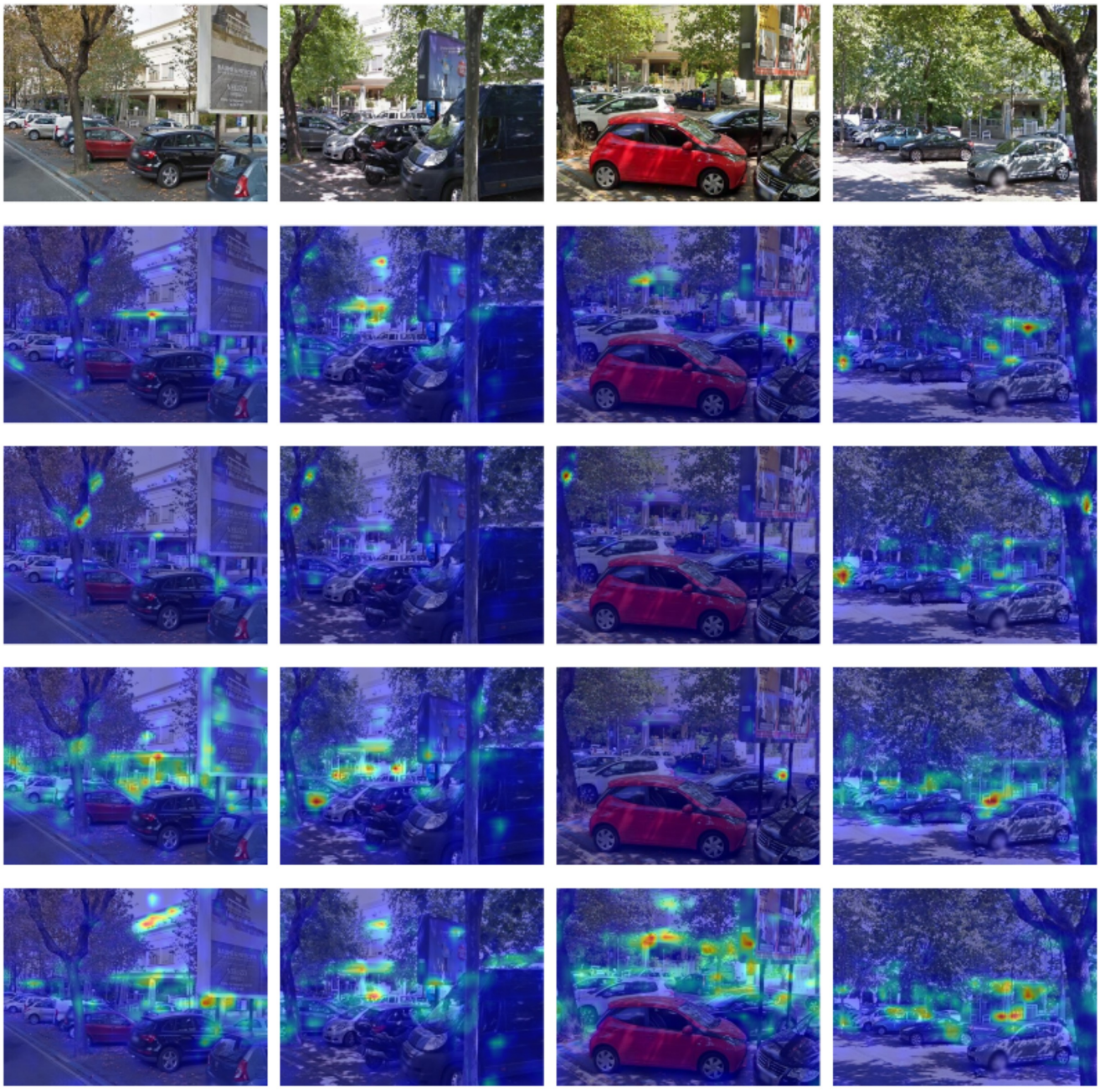

Figure 6. **Moving objects.** The consistency of attention allocation across scenes with different moving objects (cars) underscores our model's capability in distinguishing between transient and persistent features (trees, buildings) within an urban environment.

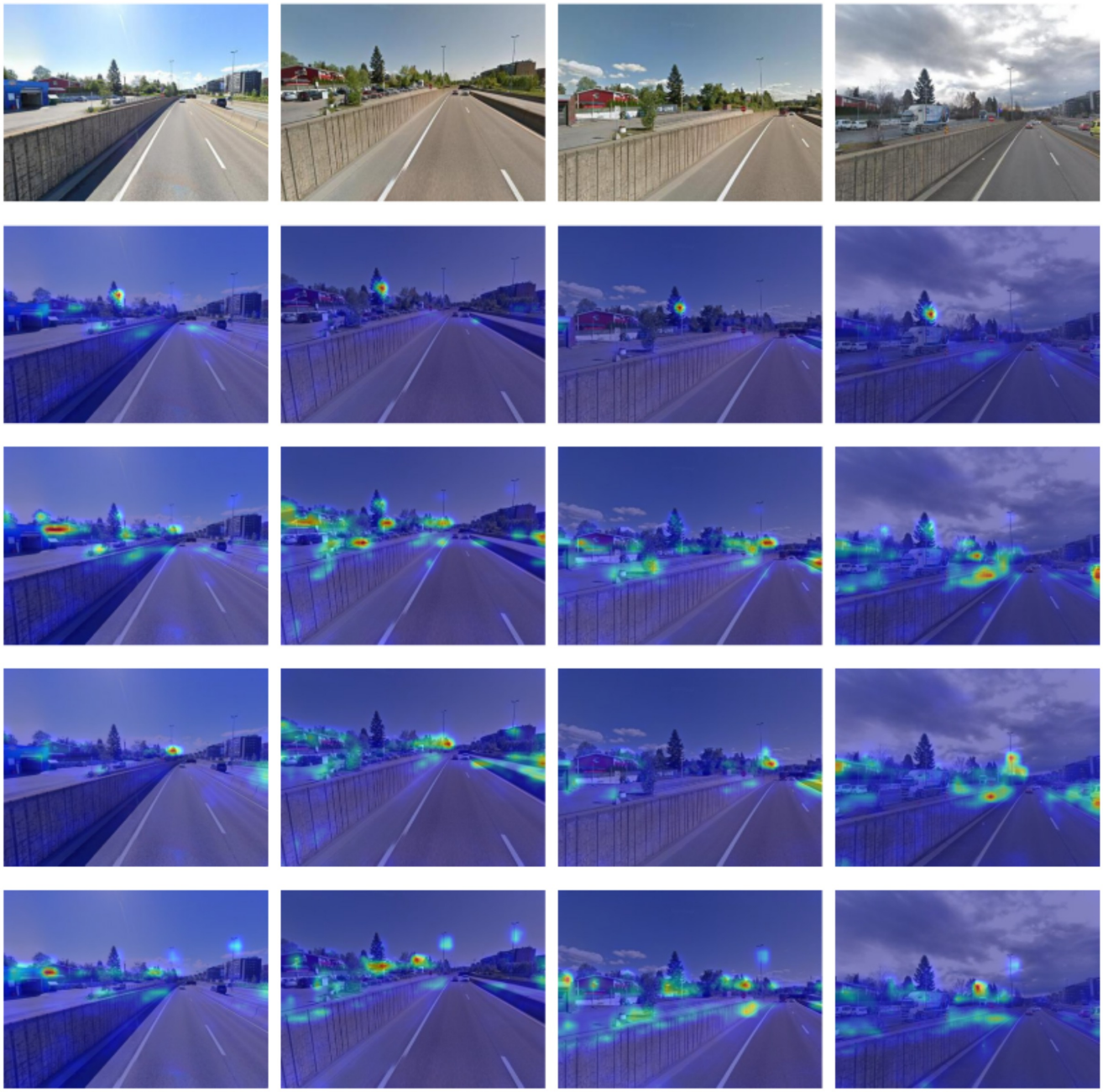

Figure 7. In this example, we can see that each query specializes in identifying particular elements within the scenes. The first query (second row) predominantly activates over big blobs of vegetation, while the second query (third row) demonstrates higher activation over architectural structures, such as buildings. These attention patterns suggest a high degree of specialization in the learned queries, enabling precise feature discrimination within complex environments.