Benchmarking the Robustness of Cross-view Geo-localization Models

Qingwang Zhang and Yingying Zhu(✉)

College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
zhangqingwang2022@email.szu.edu.cn, zhuyy@szu.edu.cn
https://zqwlearning.github.io/CrossViewRobustness

Abstract. Cross-view geo-localization serves as a viable alternative to providing geographical location information when GPS signals are unstable or unavailable by matching ground images with geo-tagged aerial image databases. While significant progress has been made on some common benchmarks like CVUSA and CVACT, there remains a lack of comprehensive consideration for robustness against real-world environmental challenges such as adverse weather or sensor noise. This deficiency poses a significant challenge for deploying this technology in safety-critical domains like autonomous driving and robot navigation. To the best of our knowledge, there is currently no specialized benchmark for evaluating the robustness of cross-view geo-localization. To comprehensively and fairly evaluate the robustness of cross-view geo-localization models in real-world scenarios, we introduce 16 common types of data corruption. By synthesizing these corruptions on public datasets, we establish two fine-grained corruption robustness benchmarks (CVUSA-C and CVACT_val-C) and three comprehensive corruption robustness benchmarks (CVUSA-C-ALL, CVACT_val-C-ALL, and CVACT_test-C-ALL), covering approximately 1.5 million corrupted images. Subsequently, we conduct large-scale experiments on various cross-view geo-localization models to evaluate their robustness in corrupted environments and derive novel insights. Finally, we explore two data augmentation strategies as potential solutions to enhance model robustness. Combined with the training strategies proposed, these approaches effectively enhance the robustness of multiple models.

Keywords: Cross-view geo-localization · Benchmarking · Robustness

1 Introduction

Geo-localization is fundamental for autonomous driving [26], robot navigation [24], 3D reconstruction [14], augmented reality [27], and road maintenance [2], *etc.* However, in the realistic scenario, a GPS-degraded or GPS-denied environment is not uncommon, including urban canyon effects, forest areas, mountainous terrain, and interference or jamming regions. The absence of geo-tags in image metadata, often resulting from GPS signal obstruction, signal attenuation, or

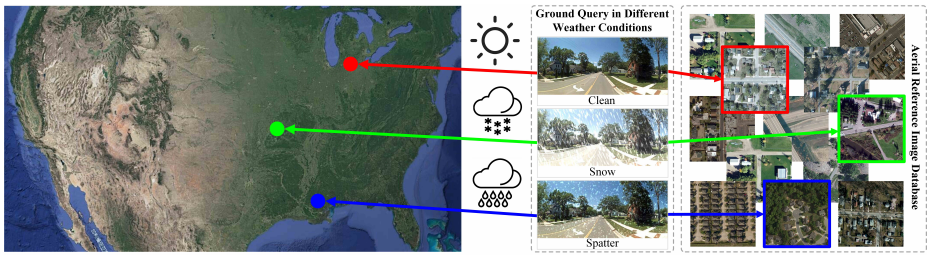


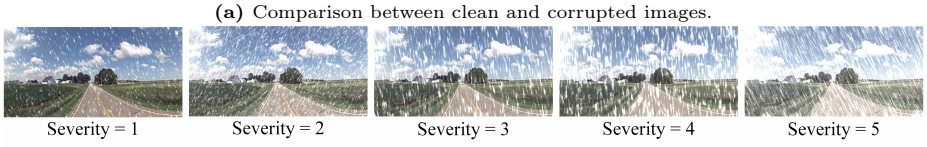
Fig. 1: Existing cross-view geo-localization models fail when the ground query image is corrupted. When the query image is clean, almost all existing methods retrieve the ground truth correctly (marked by the *red* box and the *red* circle). Once the image is corrupted, almost all existing methods have missed the correct result. Incorrect retrieve and location results are marked in *green* and *blue*.

lack of a camera localization module, leads to only 4.3% of Internet images being geo-tagged [7]. Cross-view geo-localization can determine the location where a ground-level image is taken by comparing it with a geo-tagged database of aerial images (*e.g.*, satellite images), which can serve as an alternative way or an effective complement to provide geo-location information in the case of GPS-degraded or GPS-denied. Therefore, cross-view geo-localization has recently garnered considerable attention due to its wide range of practical applications.

Two favorite benchmarks have been introduced to evaluate cross-view geo-localization performance, *e.g.*, CVUSA [40] and CVACT [21]. Existing methods are all evaluated on both datasets and state-of-the-art approaches even achieve near-perfect results [42,43,45]. Nevertheless, these benchmarks only present ideal settings and do not reflect real-world scenarios. Almost all their ground images were collected on clean days, with little coverage in conditions such as rain, snow, and sensor noise corruption. Inevitably, due to the significant differences between the distributions of the training data and the real-world data, the performance of current data-driven models degrades drastically or even fails when the models trained on "clean" data are applied to "corrupted" data. Fig. 1 shows examples of localization failures that can result from the ground-level query under different weather conditions. This poses a significant challenge to applying this task to safety-critical applications such as autonomous driving. Thus, prior to deployment, it has become imperative to create an easy-to-use corruption robustness benchmark and to comprehensively evaluate the robustness of existing models with respect to corrupted data. To the best of our knowledge, we first create cross-view geo-localization robustness benchmarks and provide an exhaustive assessment of the robustness of existing mainstream cross-view geo-localization models when faced with corrupted data.

In the cross-view geo-localization task setup, the aerial images are cached by the system in advance with high-definition and stable characteristics, while the ground-view images are collected in real-time, which is more flexible and versatile. Therefore, in this paper, we only corrupt the ground images to consider

(a) Comparison between clean and corrupted images.



(b) Comparison of different severity levels, taking snow as an example.

Fig. 2: Various types and severities of common corruption in our robustness benchmarks, after cropping are provided (best viewed when zoomed in on the screen). Visualizations of all corruption types can be found in the *supplementary material*.

the robustness of the existing models, which is more practical while alleviating the evaluation cost. We systematically design **16** common types of corruption that are commonly found in the cross-view geo-localization task and provide a comprehensive evaluation of the corruption robustness of existing models. These corruptions include **Weather**, **Blur**, **Noise**, and **Digital**, covering a wide range of real-world situations. Furthermore, each type of corruption has **5** severity levels, totaling **80** different types of corruption involved. We have applied these corruptions on CVUSA [40] and CVACT [21] to create two fine-grained corruption robustness benchmarks (CVUSA-C and CVACT_val-C) and three comprehensive corruption robustness benchmarks (CVUSA-C-ALL, CVACT_val-C-ALL, and CVACT_test-C-ALL), which cover a total of about **1.5 million** corrupted images. Fig. 2 shows some examples of corrupted images. We expect these datasets to serve as common benchmarks for comprehensively and fairly evaluating the corruption robustness of cross-view geo-localization models and to facilitate future research to advance the cross-view geo-localization task.

The contributions made to this work are summed up as follows:

- To the best of our knowledge, we have benchmarked the robustness of state-of-the-art cross-view geo-localization models against real-world data corruption challenges for the first time. We have generated about 1.5 million corrupted images based on the CVUSA and CVACT datasets to establish benchmarks for assessing the robustness of cross-view geo-localization models.
- Based on the benchmark study, we have several new insights: (1) in most cases, the clean performance (R@K_{clean}) of a model is positively, but not absolutely, correlated with its robustness; (2) snow, spatter, and zoom blur more significantly affect the robustness of various models compared to other corruptions; (3) models trained on more intricate scenarios (*e.g.*, CVACT) exhibit better robustness.
- Introducing stylization and histogram equalization as data augmentation techniques, along with our proposed training strategy, significantly enhances the robustness of various cross-view geo-localization models.

2 Related Work

2.1 Cross-view Geo-localization

Cross-view geo-localization has witnessed significant advancements in recent years. Initially, [40] introduced Convolutional Neural Networks (CNN) to the cross-view matching task, resulting in notable performance improvements. Subsequently, [17] employed a VGG [36] backbone network with two branches, combined with the NetVlad [1], and proposed the weighted soft-margin ranking loss, achieving state-of-the-art performance. To further enhance network performance, [21] emphasized the importance of orientation in cross-view geo-localization and devised a method to provide orientation information to the neural network explicitly. [33] adopted spatial-aware feature aggregation modules to aggregate information-rich and diverse feature maps, while introducing polar transformation for pre-processing to narrow the geometric gap between center-aligned satellites and ground-level images. Recently, [42] explored the integration of Vision Transformer [5] in cross-view geo-localization, proposing a novel layer-to-layer Transformer with self-cross attention mechanism that highlighted the significance of considering global dependencies to reduce visual ambiguity. [45] introduced an attention-guided pure Transformer approach, which further improved the resolution of satellite images through additional training, thus advancing the performance of the task. [43] introduced a novel geometric layout extraction module that explicitly decouples geometric information from original features and proposed two types of data augmentation methods. However, all these methods ignore the performance of the model on corrupted images, *i.e.*, robustness, which is the focus of this paper.

2.2 Robustness Benchmarks

Robustness benchmarks are essential in the field of computer vision to enhance the stability and reliability of computer vision systems when facing uncertainties and noise. For instance, IMAGENET-C and IMAGENET-P were proposed to evaluate robustness in classification tasks [16]. In the context of autonomous driving, PASCAL-C, COCO-C, and Cityscapes-C were introduced as benchmarks for evaluating the robustness of object detection tasks [13, 25]. Similarly, robustness benchmarks for semantic segmentation [23] were established using PASCAL VOC 2012 [6], Cityscapes [3], and ADE20K [18, 44]. Furthermore, KITTI-C, nuScenes-C, and Waymo-C were devised as robustness benchmarks for 3D object detection [4]. These benchmarks aid in evaluating the robustness of models under various complex environmental conditions, thus facilitating their application in the real world. However, to the best of our knowledge, no researcher has proposed relevant robustness benchmarks specifically for cross-view geo-localization. Therefore, this paper addresses this gap by introducing the benchmarks CVUSA-C, CVACT_val-C, CVUSA-C-ALL, CVACT_val-C-ALL, and CVACT_test-C-ALL, thus filling this void in the field.

2.3 Robustness Enhancement

Several strategies have been proposed to address the impact of corruption in computer vision. For example, [20] and [8] present methods that utilize patch-based priors for background and rain layers, along with a deep CNN to effectively remove rain streaks from images. In a similar vein, [15] introduced a simple yet powerful technique employing the dark channel before eliminating haze from single-input images. Moreover, [22] devised the DesnowNet, a multi-stage network specifically designed to remove snow particles from images. However, a limitation of these methods lies in their reliance on specific designs tailored to particular types of corruption. This hinders their ability to generalize and handle other types of corruption effectively. As an alternative, some approaches have sought to enhance model performance through data augmentation, incorporating corrupted data during the training process. Although fine-tuning on specific corruption data has shown promise in boosting performance for those particular corruption types [12, 39] discovered that fine-tuning on one type of corruption often struggles to generalize to other corruption types. *How to enhance robustness remains an open question.* In this work, we employ the stylization [11] and histogram equalization [28] techniques to the train set, and we observe improvements in robustness.

3 Robustness in Cross-view Geo-localization

3.1 Problem Formulation

Robustness. First, we consider a set of ground-to-aerial image pairs $\{\mathbf{I}_i^g, \mathbf{I}_i^a\}_{i=1}^N$, where N represents the number of image pairs. The superscripts g and a denote ground and aerial views, respectively. For a ground query image indexed as q, we assume the existence of ground image encoder f_g and aerial image encoder f_a, which have been trained on samples from the distribution $\mathcal{D}$. We let $\mathbb{P}_{\mathcal{C}}(c)$ approximate the frequency of real-world corruption. Existing models mostly evaluate performance when samples are drawn from the distribution $\mathcal{D}$, denoted as $\mathbb{P}_{(\mathbf{I}^g,\mathbf{I}^a)\sim\mathcal{D}}(d(f_g(\mathbf{I}_q^g), f_a(\mathbf{I}_q^a)) < \{d(f_g(\mathbf{I}_q^g), f_a(\mathbf{I}_i^a))|\forall i \in \{1, ..., N\}, i \neq q\})$, $d(\cdot,\cdot)$ representing the L_2 distance. However, in practical deployments, systems often need to operate on low-quality or corrupted images. Hence, we construct an evaluation of the corruption robustness of the model, denoted as:

$$\mathbb{E}_{c\sim\mathcal{C}}[\mathbb{P}_{(\mathbf{I}^g,\mathbf{I}^a)\sim\mathcal{D}}(d(f_g(c(\mathbf{I}_q^g)), f_a(\mathbf{I}_q^a)) < \{d(f_g(c(\mathbf{I}_q^g)), f_a(\mathbf{I}_i^a))|\forall i \in \{1, ..., N\}, i \neq q\})]$$

In this work, in order to approximate the real-world corruption $\mathcal{C}$, we designed a set of common corruptions (Weather, Blur, Noise, and Digital) in the cross-view geo-localization task. The types of corruption already provide a good overview of corruption that street view images may face. From these common corruptions, we create our cross-view geo-localization robustness evaluation benchmarks.

3.2 Image Corruption

Common corruptions are categorized based on [16], and ground-level image corruptions are divided into four major categories with a total of 16 subcategories.

The first major category is **weather-related corruptions**, which encompasses *snow*, *frost*, *fog*, *brightness*, and *spatter*. The second major category consists of **blur-related corruptions**, including *defocus*, *glass*, *motion*, and *zoom* blur. Defocus blur occurs when an image is out of focus, while glass blur arises when images are captured through frosted glass windows of sensors. Motion and zoom blurring occur in scenes with rapid camera movement or quick approaches toward objects, and they are particularly prone to occur in image collections from vehicular devices. The third major category involves **noise-related corruptions**, which cover *Gaussian*, *shot*, *impulse*, and *speckle* noise. Gaussian noise emerges under low-light conditions, attributed to the discrete nature of light and causing electronic noise. Impulse noise results from bit errors, resembling the color analog of salt-and-pepper noise. Speckle noise, which occurs due to light interference, leads to the appearance of bright and dark speckles. Lastly, the fourth category includes **digital corruptions**, including *contrast*, *pixelation*, and *JPEG* corruption. Contrast corruption is dependent on lighting conditions and the colors of the objects during capture. Pixelation occurs during upscaling low-resolution images, and JPEG corruption arises when images undergo lossy compression, introducing compression artifacts. To better simulate real-world corruption, each type of corruption also contains 5 severity levels. We present visualizations of various types and severities of common corruption in Fig. 2.

3.3 Robustness Enhancement Methods

To enhance the robustness of existing models against corruption, we investigate two data augmentation approaches: stylization and histogram equalization. By applying either stylization or histogram equalization to the training data and employing our proposed training strategy, we effectively bolstered the corruption robustness of multiple cross-view geo-localization models.

Stylization. Style transfer, introduced by [9], merges the content and style of two different images to generate a novel image that retains the content of the original image while adopting the style of the target image, which is also known as stylization. In the context of image classification and object detection tasks, style transfer has demonstrated its effectiveness in enhancing robustness [10].

Histogram Equalization. Histogram equalization is a fundamental technique used for enhancing image contrast. It operates by redistributing pixel intensity values within an image, stretching or compressing the brightness range, thus achieving a more uniform distribution of pixel intensity values across the entire brightness range of the image. In this study, we employ Contrast Limited Adaptive Histogram Equalization (CLAHE) [29] to enhance the robustness of existing methods.

Training Strategy. In this paper, we apply Stylization / CLAHE to the cross-view geo-localization training set, testing 3 settings: (1) training using the standard (raw) training data; (2) replacing all training images with Stylization / CLAHE images, thereby eliminating the use of standard images during training; (3) having Stylization / CLAHE images participate in the training process alongside the standard images with equal probability at each iteration. Fig. 3

Fig. 3: Visualization of Stylization / CLAHE applied to the CVUSA training set. The illustration depicts standard images (middle row), stylization images (top row), and CLAHE images (bottom row). The rounded rectangles on both sides represent different training strategies. The *red* dashed boxes indicate the style images sourced from `https://www.kaggle.com/c/painter-by-numbers/`. Inspired by [25].

shows examples of applying Stylization / CLAHE, and it illustrates 3 training strategies with the same training complexity.

4 Corruption Robustness Benchmarks

We create two fine-grained corruption robustness benchmarks (CVUSA-C and CVACT-C) and three comprehensive corruption robustness benchmarks (CVUSA-C-ALL, CVACT_val-C-ALL, and CVACT_test-C-ALL) from the classic cross-view geo-localization datasets CVUSA and CVACT. See Table 1 for more detailed information on these benchmarks.

4.1 Fine-grained Corruption Robustness Benchmarks

We create fine-grained robustness benchmarks comprising 16 corruption types, each corresponding to 5 severity levels denoted by integers from 1 to 5, where higher numbers indicate more severe corruption. Consequently, one single original ground query image can yield **80** (16×5) corrupted images. In the fine-grained corruption robustness benchmarks, each type of corrupted image forms an independent evaluation subset with no overlap. The suffix "-C" designates all fine-grained robustness benchmarks (CVUSA-C and CVACT_val-C).

CVUSA-C. The CVUSA [40] dataset is one of the earliest cross-view geo-localization datasets, primarily collected from suburban areas in the United States. It encompasses a total of 35,532 pairs of ground-aerial images for training and 8,884 pairs for testing. We have employed the CVUSA test set to generate the fine-grained corruption robustness benchmark, CVUSA-C. To be specific, CVUSA-C comprises all types and severities of image corruptions, each residing independently within individual subsets of CVUSA-C. Given that there are 16 distinct corruption types and 5 levels of corruption severity, CVUSA-C effectively comprises **80** evaluation subsets. Each of these subsets contains 8,884 ground images for testing, resulting in an aggregate of 710,720 corrupted images.

Table 1: Detailed information on the proposed corruption robustness benchmarks.

	Fine-grained		Comprehensive		
	CVUSA-C	CVACT_val-C	CVUSA-C-ALL	CVACT_val-C-ALL	CVACT_test-C-ALL
Number of original validation / test ground images	8,884	8,884	8,884	8,884	92,802
Whether or not evaluation subsets are generated for each corruption	✔	✔	✘	✘	✘
Whether all corruptions are included in an independent subset	✘	✘	✔	✔	✔
Number of validation / test ground images for our benchmark	8,884 × 16 × 5	8,884 × 16 × 5	8,884	8,884	92,802
Storage space	∼ 39 GB	∼ 178 GB	∼ 0.5 GB	∼ 2 GB	∼ 21 GB

CVACT_val-C. The CVACT [21] dataset was collected by [21] and densely covers an urban area of Australia (Canberra). Similar to the CVUSA dataset, this dataset comprises 35,532 pairs of ground-aerial images for training and 8,884 image pairs for validation (referred to as CVACT_val). We create a fine-grained robustness benchmark from CVACT_val, as CVUSA does. The benchmark includes various types and severities, resulting in a total of **80** validation subsets. These subsets consist of 710,720 corrupted images and are collectively referred to as CVACT_val-C.

Evaluation Metrics. The original evaluation metrics for cross-view geo-localization primarily revolved around R@K ($K \in \{1, 5, 10, 1\%\}$), which represents the probability of correctly identifying the matching image within the top K retrieved reference images based on the query image [17, 21, 33, 42, 45]. We denote the performance of models on the original validation or test sets as R@K_{clean}. For each corruption type c and each severity level s, we employ R@$K_{c,s}$ to gauge the performance of models under corruption conditions. Finally, by averaging across all corruption types and severity levels, we compute the average performance of model R@K_{cor} under corruption conditions. In this context, $\mathcal{C}$ represents the set of corruptions in evaluation. Additionally, we calculate the Relative Corruption Error (RCE) by measuring the percentage of performance degradation [4]. A higher RCE implies poorer robustness.

$$\text{R@}K_{\text{cor}} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\frac{1}{5}\sum_{s=1}^{5}\text{R@}K_{c,s} \tag{1}$$

$$\text{RCE}_{c,s} = \frac{\text{R@}K_{\text{clean}} - \text{R@}K_{c,s}}{\text{R@}K_{\text{clean}}}; \text{RCE} = \frac{\text{R@}K_{\text{clean}} - \text{R@}K_{\text{cor}}}{\text{R@}K_{\text{clean}}} \tag{2}$$

4.2 Comprehensive Corruption Robustness Benchmarks

Furthermore, we introduce comprehensive corruption robustness benchmarks. Prior efforts [4, 16, 25] often generated separate test sets for each corruption type and its corresponding 5 severity levels. While this approach allows for a finer-grained evaluation of each corruption type's impact on the model, it imposes significant storage and computational costs (**80** evaluation subsets). Recognizing that real-world concerns may prioritize a model's overall performance against diverse corruptions, for each corruption type and severity level, *we aggregate all forms of corruption into a single evaluation set*, creating comprehensive benchmarks. The suffix "-C-ALL" designates all comprehensive corruption robustness benchmarks (CVUSA-C-ALL, CVACT_val-C-ALL, and CVACT_test-C-ALL).

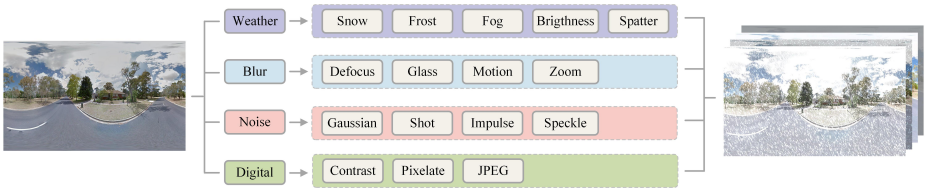


Fig. 4: The proposed fine-grained and comprehensive robustness benchmarks. Each corruption category encompasses 5 severity levels. The primary distinction between fine-grained and comprehensive robustness benchmarks lies in whether separate evaluation subsets are created for each corruption category and severity level.

CVUSA-C-ALL and CVACT_val-C-ALL. For the 8,884-pair test set of the CVUSA dataset and the 8,884-pair validation set of the CVACT_val dataset, we designed comprehensive corruption robustness benchmarks. These benchmarks are denoted as CVUSA-C-ALL and CVACT_val-C-ALL, respectively. Each dataset contains 8,884 corrupted images and systematically covers all corruption types and severity levels.

CVACT_test-C-ALL. The CVACT dataset additionally provides 92,802 pairs of images for testing purposes (referred to as CVACT_test), showcasing dense coverage of urban areas. Similar to CVUSA-C-ALL and CVACT_val-C-ALL, we create a comprehensive corruption robustness benchmark from CVACT_test, named CVACT_test-C-ALL, containing a total of 92,802 corrupted images, covering all corruption types and severity levels.

Evaluation Metrics. Since the comprehensive corruption robustness benchmarks represent independent subsets for testing, the recall accuracy (R@K_{all}, $K_{\text{all}} \in \{1, 5, 10, 1\%\}$) can be directly employed for evaluation.

The comprehensive corruption robustness benchmarks serve as valuable tools for evaluating the overall robustness of models under all corruption scenarios, and they effectively reduce the cost of robustness evaluation.

In summary, fine-grained robustness benchmarks provide a more detailed performance evaluation, while comprehensive robustness benchmarks provide an overall performance evaluation and reduce the evaluation cost. An illustrative diagram of the corruption robustness benchmarks is presented in Fig. 4.

5 Benchmarking Results

We have selected classical cross-view geo-localization models to benchmark on the proposed benchmarks. The experimental results on fine-grained corruption robustness benchmarks are shown in Section 5.1. Section 5.2 presents the experimental results on comprehensive corruption robustness benchmarks. In Section 5.3, we report the experimental results demonstrating the influence of stylization and histogram equalization on the robustness of models. *Supplementary material* provides additional experimental results.

Table 2: Experimental results of 8 cross-view geo-localization methods on the CVUSA-C benchmark. We report the R@1 performance of each method under different corruptions (obtained by averaging the 5 corruption severities), as well as the average $R@1_{cor}$ under all corruption types.

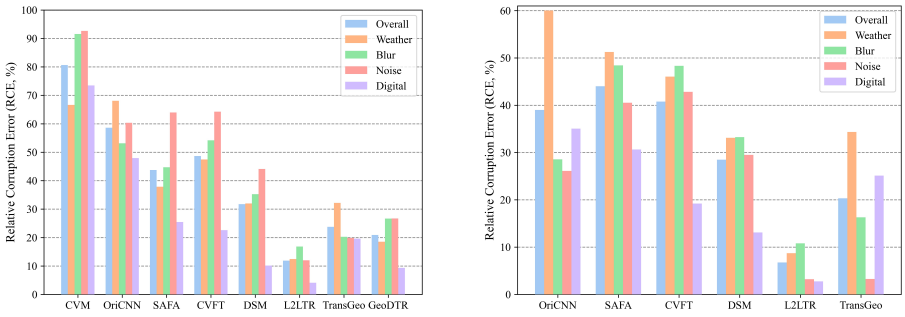
Method	Clean	CVUSA-C																$R@1_{cor}$
		Weather					Blur				Noise				Digital			
		Snow	Frost	Fog	Bright	Spatter	Defocus	Glass	Motion	Zoom	Gaussian	Shot	Impulse	Speckle	Contrast	Pixel	JPEG	
CVM-Net [17]	22.47	0.86	8.42	8.37	13.75	6.11	1.06	4.81	1.47	0.23	1.82	1.18	1.28	2.32	4.75	6.89	6.23	4.35
OriCNN [21]	40.79	7.36	6.51	7.57	21.69	22.01	20.46	26.10	19.60	10.32	17.24	13.95	19.40	14.09	7.94	28.51	27.27	16.88
SAFA [33]	89.84	19.32	60.42	67.63	81.96	49.86	51.24	80.56	55.49	11.44	33.04	28.51	30.37	37.59	31.67	88.05	81.15	50.52
CVFT [35]	61.43	8.00	30.79	47.46	47.54	27.63	24.55	44.93	34.89	8.17	21.83	19.19	20.56	26.25	38.28	57.25	47.11	31.53
DSM [34]	91.96	24.24	64.44	84.08	82.44	57.58	62.48	84.52	66.02	25.15	49.55	46.40	48.84	60.83	72.11	90.20	85.56	62.78
L2LTR [42]	94.05	**67.19**	**85.00**	92.64	91.61	**75.24**	**88.35**	93.13	**89.33**	**42.07**	**81.32**	**80.29**	**82.88**	**86.54**	**86.36**	93.64	**90.56**	**82.88**
TransGeo [45]	94.08	29.39	69.50	70.89	85.01	64.26	80.97	92.16	85.96	40.97	72.95	70.27	74.32	83.99	43.01	93.74	90.13	71.72
GeoDTR [43]	**95.43**	44.20	84.95	**92.80**	**93.55**	73.14	82.64	**93.29**	76.80	27.19	68.40	64.45	68.53	78.28	74.80	**94.45**	90.20	75.48

5.1 Fine-grained Corruption Robustness Benchmarking Results

Benchmarking Results on CVUSA-C Table 2 presents the results of 8 cross-view geo-localization methods on the fine-grained corruption robustness benchmark CVUSA-C. In general, the robustness of the model is positively correlated with its accuracy on clean data. These models with higher $R@1_{clean}$, such as L2LTR [42], TransGeo [45], and GeoDTR [43], also achieve higher $R@1_{cor}$, which is understandable as different models show consistent performance degradation on corrupted images. It's worth noting that L2LTR outperforms TransGeo and GeoDTR in $R@1_{cor}$, even though TransGeo and GeoDTR exhibit higher $R@1_{clean}$. This observation differs from the results obtained in [18] and [4]. The discrepancy might be attributed to its deeper network architecture and more parameters, which make it more robust to input variations. This finding emphasizes the importance of considering the robustness of the model independently, especially when its performance on clean data is high, as *higher clean performance does not necessarily indicate stronger robustness*. In Fig. 5a, we plot the RCE of models under various corruption types (according to major categories). Based on our experimental results, we derive the following insights.

Impact of Corruption Types. Based on the results from Table 2 and Fig. 5a, we observed that snow, spatter, and zoom corruptions have a significant impact on the performance of various cross-view geo-localization models in the CVUSA-C benchmark, resulting in RCE values exceeding 28% for all models. Conversely, the effects of glass blur, JPEG, and pixelation on performance are relatively minor. These findings demonstrate the threats posed by adverse weather conditions and fast motion to cross-view geo-localization models. In contrast, most models exhibited less performance degradation under the influence of glass blur, JPEG compression, and pixelation, possibly due to similar corruptions being present in the training dataset, allowing the models to learn prior knowledge about these types of corruptions.

Performance of Different Models. Among all the evaluated models, L2LTR [42] exhibited the most outstanding $R@1_{cor}$ performance. Additionally, we observed a synchronized growth trend between $R@1_{clean}$ and $R@1_{cor}$ for all methods except L2LTR. Notably, CVM-Net [17] demonstrated the weakest corruption robustness. All of the methods show an average performance degradation



(a) CVUSA-C **(b)** CVACT_val-C

Fig. 5: The Relative Corruption Error (RCE) of different cross-view geo-localization models on CVUSA-C and CVACT_val-C benchmarks. The results are shown for each major category of corruptions, including **Weather**, **Blur**, **Noise**, and **Digital**, as well as the overall performance across all types of corruptions (**Overall**).

Table 3: Experimental results of 7 cross-view geo-localization methods on the CVACT_val-C benchmark. We report the R@1 performance of each method under different corruptions (obtained by averaging the 5 corruption severities), as well as the average $R@1_{cor}$ under all corruption types.

Method	Clean	CVACT_val-C																
		Weather					Blur				Noise				Digital			$R@1_{cor}$
		Snow	Frost	Fog	Bright	Spatter	Defocus	Glass	Motion	Zoom	Gaussian	Shot	Impulse	Speckle	Contrast	Pixel	JPEG	
OriCNN [21]	46.96	13.94	6.13	3.78	29.45	40.54	31.71	39.99	37.58	24.89	34.24	32.27	39.01	33.28	4.56	44.38	42.56	28.65
SAFA [33]	81.03	20.03	31.66	33.19	66.99	45.60	39.83	72.87	49.86	4.62	48.66	43.68	48.82	51.61	15.91	76.90	75.83	45.38
CVFT [35]	61.05	15.00	22.32	42.53	47.60	37.25	31.30	53.88	36.91	4.10	35.68	30.80	36.32	36.84	31.79	58.21	57.97	36.16
DSM [34]	82.49	31.95	51.70	70.43	69.48	52.35	57.35	80.16	67.38	15.34	58.34	53.05	58.18	63.06	52.79	81.72	80.55	58.99
L2LTR [42]	84.89	**71.03**	**77.93**	**83.50**	81.17	**73.78**	**83.98**	85.07	**84.00**	**49.79**	**82.20**	**81.19**	**82.98**	82.23	**79.15**	85.07	83.40	**79.15**
TransGeo [45]	84.95	47.65	58.51	32.91	72.67	67.13	81.43	84.83	81.80	36.34	81.96	80.86	82.84	**83.01**	22.18	84.92	83.74	67.68
GeoDTR [43]	**86.21**	48.24	71.74	83.26	**84.60**	61.39	79.11	**85.51**	73.44	8.26	75.44	73.99	77.06	80.23	55.48	**86.01**	**85.19**	70.56

of **24.24%** on the CVUSA-C benchmark, with SAFA [33] degrading by up to 39.32% and L2LTR [42] degrading by a minimum of 11.17%.

Benchmarking Results on CVACT_val-C Table 3 presents the results of 7 cross-view geo-localization methods on the fine-grained corruption robustness benchmark CVACT_val-C. Fig. 5b shows the RCE of models across various corruption types. Similar to the experimental results for the CVUSA-C benchmark, L2LTR [42] exhibits the best robustness. Meanwhile, snow, spatter, and zoom corruptions hurt the performance of the various models the most, and the various models have better robustness to glass blur, JPEG, and pixelation corruptions. However, it is worth noting that all methods show an average performance degradation of **20.14%** on the CVACT_val-C benchmark.

Effects of different training sets. From the average results of all methods of CVUSA-C and CVACT_val-C benchmarks, the $R@1_{clean}$ to $R@1_{cor}$ drop of CVACT_val-C benchmark is 4.1% less than CVUSA-C. We believe that this may be due to the fact that the CVACT dataset is collected in urban areas, while the CVUSA dataset is collected in the suburbs, and thus the CVACT dataset has *richer scene information and so brings more robustness to the model.*

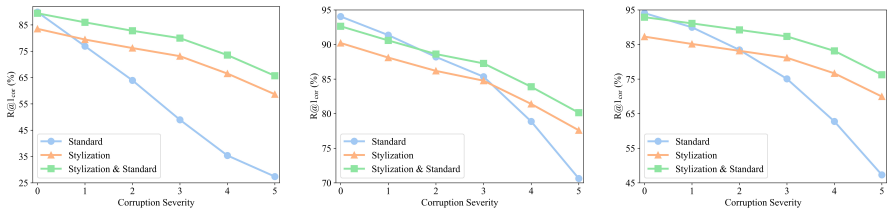
Table 4: Experimental results of cross-view geo-localization methods on CVUSA-C-ALL, CVACT_val-C-ALL, and CVACT_test-C-ALL benchmarks.

Method	CVUSA-C-ALL				CVACT_val-C-ALL				CVACT_test-C-ALL			
	$R@1_{all}$	$R@5_{all}$	$R@10_{all}$	$R@1\%_{all}$	$R@1_{all}$	$R@5_{all}$	$R@10_{all}$	$R@1\%_{all}$	$R@1_{all}$	$R@5_{all}$	$R@10_{all}$	$R@1\%_{all}$
CVM-Net [17]	6.09	16.05	23.14	52.51	-	-	-	-	-	-	-	-
OriCNN [21]	9.38	22.26	30.04	58.99	15.31	28.31	35.21	58.39	3.69	8.33	11.04	43.93
SAFA [33]	63.68	78.08	82.82	93.91	56.72	73.60	78.59	91.32	31.18	52.06	58.60	90.41
CVFT [35]	41.05	64.01	72.64	91.37	45.69	66.45	72.97	88.38	22.82	43.48	51.07	88.99
DSM [34]	75.27	86.26	89.42	95.07	70.04	82.81	85.86	93.51	47.13	68.41	73.52	93.18
L2LTR [42]	**87.93**	**95.45**	**97.01**	**99.01**	**82.13**	**93.34**	**94.93**	**98.10**	**57.20**	**82.59**	**87.23**	**98.09**
TransGeo [45]	82.72	91.95	94.03	97.92	74.04	86.19	89.10	94.98	52.18	74.35	78.99	95.03
GeoDTR [43]	84.64	93.29	95.01	98.24	77.40	88.95	91.28	95.91	52.87	78.84	83.17	95.84

5.2 Comprehensive Corruption Robustness Benchmarking Results

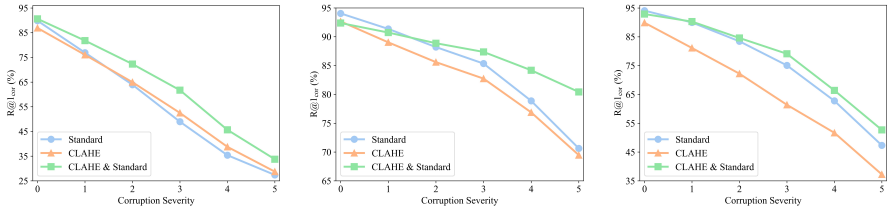
The performances of different models on comprehensive corruption robustness benchmarks (CVUSA-C-ALL, CVACT_val-C-ALL, and CVACT_test-C-ALL) are shown in Table 4. It is clear from the experimental results that the performance degradation is closely and positively correlated with the original performance when evaluated using the comprehensive corruption robustness benchmarks, except for L2LTR [42]. The L2LTR exhibits the highest level of robustness, albeit at the expense of increased computational cost and a greater number of trainable parameters. The results obtained by all methods for the comprehensive robustness benchmarks follow the same trend as the $R@1_{cor}$ results for the fine-grained robustness benchmarks. This demonstrates the benefit of the comprehensive robustness benchmarks when considering overall robustness; they allow the best robustness model to be selected with only a small evaluation cost (as in the normal evaluation process).

5.3 Robustness Enhancement Results



(a) SAFA **(b)** L2LTR **(c)** TransGeo

Fig. 6: Stylization improves the robustness of SAFA [33], L2LTR [42], and TransGeo [45] on the CVUSA-C benchmark, with each severity level representing the average across all 16 corruption types. Severity = 0 corresponds to clean images for testing. The Standard denotes the original, unaltered training data, while Stylization denotes training exclusively on images subjected to stylization. Stylization & Standard denotes stylization and origin training data are equally interleaved during the training process. Notably, the three different training strategies require identical training complexity, and experimental configurations and model structures remain consistent throughout.



(a) SAFA **(b)** L2LTR **(c)** TransGeo

Fig. 7: CLAHE improves the robustness of SAFA [33], L2LTR [42], and TransGeo [45] on the CVUSA-C benchmark, with each severity level representing the average across all 16 corruption types. Notations are similar to Fig. 6.

Stylization for Robustness Enhancement We investigated whether the use of stylization enhances the robustness of various cross-view geo-localization models. We evaluated the performance of 3 classical cross-view geo-localization models on the CVUSA-C benchmark using 3 different training strategies, as depicted in Section 3.3 and Fig. 3. Fig. 6 shows the experimental results.

Training on stylized images indeed resulted in stronger robustness compared to models trained only on clean images, showing less performance degradation as the severity of corruption increased. However, the performance on the original clean images (severity = 0) is noticeably poorer. This can be attributed to the fact that stylized data alters the distribution of the original data. By equally combining stylized and clean data during training (Stylization & Standard), a trade-off between clean and corrupted performance was achieved, leading to both high performance akin to the standard data on clean instances and significantly improved performance on corrupted data. Results from Fig. 6 indicate that both Stylization and Stylization & Standard training enhanced $R@1_{cor}$ under corruption, especially when corruption level is greater (*e.g.*, Severity = 4 / 5).

Histogram Equalization for Robustness Enhancement We have further examined the impact of employing histogram equalization on the robustness of cross-view geo-localization models. In our study, we employ Contrast Limited Adaptive Histogram Equalization (CLAHE) [29] to enhance the robustness of existing methods. Fig. 7 shows the experimental results.

It is evident that training models only on data that has undergone CLAHE does not greatly improve their robustness. On the other hand, our training strategy 3, which combines CLAHE and clean data in equal proportions, can somewhat increase the robustness of these models, though the improvement is less than that of stylization-based methods. It's also noteworthy that different models respond differently to the same training set of data. This emphasizes the importance of model robustness and the challenge of attracting researchers to invest in model robustness.

We must emphasize that how to improve model robustness is still an open topic, which requires the efforts of more researchers.

6 Further Discussion

Real-world corruption can be much more extensive and complex. Although we cannot enumerate all possible real-world corruptions, we have systematically designed 16 corruption types, each with 5 severity levels. These types can provide a large amount of rich data to support machine learning and computer vision research. Compared to images that must be captured in the real world, synthetic imagery can provide highly controllable and diverse data for research by adjusting weather conditions, severity, and scene types as needed. At the same time, it circumvents the security risks that may be associated with capturing images of real scenes under adverse weather conditions, as well as the privacy and ethical issues that may be involved in capturing in specific environments. Therefore, we strongly believe that our benchmarks can serve as *a practical testbed to perform controllable robustness evaluation*, which in turn will advance research on cross-view geo-localization robustness.

Our focus is on evaluating how well cross-view geo-localization models withstand the corruption of input images. However, the corruption robustness benchmarks we present can be applied to more research topics. For example, robustness evaluations and enhancements on cross-view image synthesis [30, 31, 37, 38], cross-view camera pose estimation [19, 32, 41] and autonomous driving [14, 26] can also benefit from our benchmarks.

7 Conclusion

This paper systematically investigates the impact of corruption data on cross-view geo-localization models, which is a challenge previously overlooked in the context of cross-view geo-localization studies. In contrast to "clean" data, "corrupted" data is more common and closer to the realistic world, and exhaustive experiments have shown these models invariably experience significant performance degradation in the face of corrupted data. This further validates the necessity and importance of robustness benchmarking for cross-view geo-localization. To address this crucial issue and track future developments, we propose two fine-grained corruption robustness benchmarks (CVUSA-C and CVACT_val-C) and three comprehensive corruption robustness benchmarks (CVUSA-C-ALL, CVACT_val-C-ALL, and CVACT_test-C-ALL) for the cross-view geo-localization task. Extensive experiments are conducted to evaluate existing classical methods on these corruption robustness benchmarks, revealing new insights. Furthermore, we introduce two simple techniques (stylization and histogram equalization) and the training strategy to effectively enhance robustness, without requiring any model adjustments or introducing additional training complexity. We hope that the corruption robustness benchmarks, in-depth analysis, and insightful findings presented in this paper will raise awareness within the community regarding the robustness of cross-view geo-localization models and contribute to enhancing their robustness to challenges in complex real-world environments in the future.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62072318 and Grant U22A2079, in part by the Key Project of Department of Education of Guangdong Province under Grant 2023ZDZX1016, and in part by Shenzhen Science and Technology Program under Grant 20220810142553001.

References

1. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5297–5307 (2016)
2. Chaabane, M., Gueguen, L., Trabelsi, A., Beveridge, R., O'Hara, S.: End-to-end learning improves static object geo-localization from video. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2063–2072 (2021)
3. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
4. Dong, Y., Kang, C., Zhang, J., Zhu, Z., Wang, Y., Yang, X., Su, H., Wei, X., Zhu, J.: Benchmarking robustness of 3d object detection to common corruptions in autonomous driving. arXiv preprint arXiv:2303.11040 (2023)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
6. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**, 303–338 (2010)
7. Friedland, G., Vinyals, O., Darrell, T.: Multimodal location estimation. In: Proceedings of the 18th ACM international conference on Multimedia. pp. 1245–1252 (2010)
8. Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., Paisley, J.: Removing rain from single images via a deep detail network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3855–3863 (2017)
9. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)
10. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018)
11. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. International Conference on Learning Representations,International Conference on Learning Representations (Sep 2018)

12. Geirhos, R., Temme, C.R., Rauber, J., Schütt, H.H., Bethge, M., Wichmann, F.A.: Generalisation in humans and deep neural networks. Advances in neural information processing systems **31** (2018)
13. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014)
14. Häne, C., Heng, L., Lee, G.H., Fraundorfer, F., Furgale, P., Sattler, T., Pollefeys, M.: 3d visual perception for self-driving cars using a multi-camera system: Calibration, mapping, localization, and obstacle detection. Image and Vision Computing **68**, 14–27 (2017)
15. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE transactions on pattern analysis and machine intelligence **33**(12), 2341–2353 (2010)
16. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019)
17. Hu, S., Feng, M., Nguyen, R.M., Lee, G.H.: Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7258–7267 (2018)
18. Kamann, C., Rother, C.: Benchmarking the robustness of semantic segmentation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8828–8838 (2020)
19. Lentsch, T., Xia, Z., Caesar, H., Kooij, J.F.: Slicematch: Geometry-guided aggregation for cross-view pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17225–17234 (2023)
20. Li, Y., Tan, R.T., Guo, X., Lu, J., Brown, M.S.: Rain streak removal using layer priors. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2736–2744 (2016)
21. Liu, L., Li, H.: Lending orientation to neural networks for cross-view geo-localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5624–5633 (2019)
22. Liu, Y.F., Jaw, D.W., Huang, S.C., Hwang, J.N.: Desnownet: Context-aware deep network for snow removal. IEEE Transactions on Image Processing **27**(6), 3064–3073 (2018)
23. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
24. McManus, C., Churchill, W., Maddern, W., Stewart, A.D., Newman, P.: Shady dealings: Robust, long-term visual localisation using illumination invariance. In: 2014 IEEE international conference on robotics and automation (ICRA). pp. 901–906. IEEE (2014)
25. Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A.S., Bethge, M., Brendel, W.: Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv preprint arXiv:1907.07484 (2019)
26. Middelberg, S., Sattler, T., Untzelmann, O., Kobbelt, L.: Scalable 6-dof localization on mobile devices. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13. pp. 268–283. Springer (2014)
27. Mithun, N.C., Minhas, K.S., Chiu, H.P., Oskiper, T., Sizintsev, M., Samarasekera, S., Kumar, R.: Cross-view visual geo-localization for outdoor augmented reality. In: 2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR). pp. 493–502. IEEE (2023)

28. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Computer Vision, Graphics, and Image Processing p. 355–368 (Aug 1987). https://doi.org/10.1016/s0734-189x(87)80186-x, http://dx.doi.org/10.1016/s0734-189x(87)80186-x
29. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Computer vision, graphics, and image processing **39**(3), 355–368 (1987)
30. Regmi, K., Borji, A.: Cross-view image synthesis using conditional gans. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 3501–3510 (2018)
31. Shi, Y., Campbell, D., Yu, X., Li, H.: Geometry-guided street-view panorama synthesis from satellite imagery. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 10009–10022 (2022)
32. Shi, Y., Li, H.: Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17010–17020 (2022)
33. Shi, Y., Liu, L., Yu, X., Li, H.: Spatial-aware feature aggregation for image based cross-view geo-localization. Advances in Neural Information Processing Systems **32** (2019)
34. Shi, Y., Yu, X., Campbell, D., Li, H.: Where am i looking at? joint location and orientation estimation by cross-view matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4064–4072 (2020)
35. Shi, Y., Yu, X., Liu, L., Zhang, T., Li, H.: Optimal feature transport for cross-view image geo-localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11990–11997 (2020)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (Jan 2015)
37. Tang, H., Xu, D., Yan, Y., Torr, P.H., Sebe, N.: Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7870–7879 (2020)
38. Toker, A., Zhou, Q., Maximov, M., Leal-Taixé, L.: Coming down to earth: Satellite-to-street view synthesis for geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6488–6497 (2021)
39. Vasiljevic, I., Chakrabarti, A., Shakhnarovich, G.: Examining the impact of blur on recognition by convolutional networks. arXiv preprint arXiv:1611.05760 (2016)
40. Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3961–3969 (2015)
41. Xia, Z., Booij, O., Manfredi, M., Kooij, J.F.: Visual cross-view metric localization with dense uncertainty estimates. In: European Conference on Computer Vision. pp. 90–106. Springer (2022)
42. Yang, H., Lu, X., Zhu, Y.: Cross-view geo-localization with layer-to-layer transformer. Advances in Neural Information Processing Systems **34**, 29009–29020 (2021)
43. Zhang, X., Li, X., Sultani, W., Zhou, Y., Wshah, S.: Cross-view geo-localization via learning disentangled geometric layout correspondence. arXiv preprint arXiv:2212.04074 (2022)

44. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**, 302–321 (2019)
45. Zhu, S., Shah, M., Chen, C.: Transgeo: Transformer is all you need for cross-view image geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1162–1171 (2022)