
# IMAGE-TO-IMAGE TRANSLATION for ENHANCED FEATURE MATCHING, IMAGE RETRIEVAL and VISUAL LOCALIZATION

# IMAGE-TO-IMAGE TRANSLATION FOR ENHANCED FEATURE MATCHING, IMAGE RETRIEVAL AND VISUAL LOCALIZATION

Markus S. Mueller[1,*], Torsten Sattler[2], Marc Pollefeys[3,4], Boris Jutzi[1]

[1] Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Germany
(markus.mueller5, boris.jutzi)@kit.edu
[2] Department of Electrical Engineering, Chalmers University of Technology, Sweden - torsat@chalmers.se
[3] Department of Computer Science, ETH Zurich, Switzerland - marc.pollefeys@inf.ethz.ch
[4] Microsoft

**KEY WORDS:** Image-to-Image Translation, Convolutional Neural Networks, Generative Adversarial Networks, Data Augmentation, 3D Models, Feature Matching, Image Retrieval, Visual Localization

**ABSTRACT:**

The performance of machine learning and deep learning algorithms for image analysis depends significantly on the quantity and quality of the training data. The generation of annotated training data is often costly, time-consuming and laborious. Data augmentation is a powerful option to overcome these drawbacks. Therefore, we augment training data by rendering images with arbitrary poses from 3D models to increase the quantity of training images. These training images usually show artifacts and are of limited use for advanced image analysis. Therefore, we propose to use image-to-image translation to transform images from a *rendered* domain to a *captured* domain. We show that translated images in the *captured* domain are of higher quality than the rendered images. Moreover, we demonstrate that image-to-image translation based on rendered 3D models enhances the performance of common computer vision tasks, namely feature matching, image retrieval and visual localization. The experimental results clearly show the enhancement on translated images over rendered images for all investigated tasks. In addition to this, we present the advantages utilizing translated images over exclusively captured images for visual localization.

## 1. INTRODUCTION

The performance of common machine learning algorithms typically scales with the quantity and quality of training data utilized to optimize them. Deep learning with Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) pushed the performance of learning based approaches in the recent years. Therefore, the demand for training data increased and training data sets for numerous tasks were recently published. In this contribution, we generate new training images by image-to-image translation to subsequently improve performance of common computer vision and photogrammetry tasks. We will refer to image-to-image translation as image translation for simplicity reasons in this work.

The general term for generating new training samples to enlarge data sets is widely known as data augmentation. Augmenting training data is a powerful option to overcome challenges in several fields of computer vision, like feature matching, image retrieval and visual localization. Such data augmentation includes the modification of existing training images as well as the generation of new images to expand training sets. Common methods in image processing are to shift, rotate, scale, flip, crop, transform, compress or blur training images to extend a basis data set. In this contribution new images are rendered and furthermore translated by a GAN to augment a data set of images. CNNs and other learning based methods benefit from a variety of training data. If more variety of training samples is considered in a training set, more robust and accurate networks can be expected.

*Image Translation* made a huge leap in recent years benefiting from uprising deep learning algorithms and a better understanding of such. The aim of image translation is to translate images from one domain into another, like translations between daytime and nighttime, translations between the four seasons spring, summer, autumn and winter or even the translation of artistically styles. Our goal of this contribution is to investigate the feasibility of image translation to improve computer vision tasks. The success of training based algorithms like deep learning or image retrieval depends highly on the provided training data and suffers from deficient training data. An insufficient variety of training images weakens the estimates in terms of robustness or accuracy. Therefore, we focus on expanding training sets to increase the performance of training based algorithms. Typical training data for image retrieval or visual localization consist of images captured in a specific environment, their related poses and optionally intrinsic camera calibration parameters. Augmenting such training sets to generate a higher quantity and variety of training samples has the potential to enhance methods that learn from this data. An augmentation could be undertaken by capturing additional images manually, determining their poses and adding them to an existing training set. However, in this contribution we augment existing training data with synthetic images. This is carried out by generating additional training images to a provided training set. These additional images are generated by utilizing only the pre-existing captured data of a benchmark data set. There is no necessity for further assumptions or manual capturing of new data. Given an image data set consisting of images and their corresponding poses of a specific environment we generate a 3D model of the scene by utilizing a Structure-from-Motion (SfM) pipeline. Images with arbitrary poses are rendered in this model. These images are used to enhance the training data set. However, the rendered images differ strongly in appearance from the original captured training images since the 3D model

*Corresponding author

is no photo-realistic representation of the scene. The generation of photo-realistic models is challenging and not yet fully automatized. Hence, we create a simple triangulated model of the environment. Since the straight utilization of such rendered images may not suffice as training data for further applications, we apply image translation. By image translation we transform the rendered images from their *rendered* domain into a more realistic domain, namely the *captured* domain. Therefore, the translated images have a higher similarity to the originally captured images. This higher similarity to the original training images increases the feasibility for potentially serving as additional training data. For image translation again there is no need to capture new data nor to make additional assumptions. The image translation pipeline is trained only on the original training set and the rendered images. The rendered images are generated from the 3D model, which is again created by only utilizing the original training images. Therefore, we combine image rendering with image translation for data augmentation to enhance common computer vision tasks. The evaluation of the newly generated training data, namely the images translated from the *rendered* domain into the *captured* domain, is carried out by performing common computer vision tasks on them. In detail, we perform feature matching, image retrieval and visual localization to investigate the beneficial impact of image translation.

*Feature Matching* is a fundamental algorithm for image analysis. Local features are extracted and characterized by their descriptors. These descriptors can be compared and matched according to their similarity. A lot of computer vision tasks utilize feature matching, e.g. classification, segmentation, detection, image retrieval, 3D reconstruction, tracking methods or image alignment. Image translation has the potential to enhance this fundamental algorithm by transforming images from different domains into one concurrent domain. The images radiometry is transformed, whereas the mutual similarity of them increases. Therewith, one can suppose that the similarity of extracted features of these images also increases and in turn enhances feature matching.

*Image Retrieval* is the task of finding the most similar image in a set of images given a query image. One of such image retrieval methods is Content Based Image Retrieval (CBIR), where colors, shapes or textures of an image are analyzed by computer vision algorithms to find similarities between two or more images. In our case we extract features followed by histogram intersection to find the most similar training images given a query image. By providing poses for the training images, a pose for a test image can be determined by simply assigning the pose of its nearest neighbour or more complex variants like a weighted pose of multiple nearest neighbours. As depicted above, we aim to extend such training sets by rendered and translated images to increase the provided number of images and poses. An increased number of poses and a denser distribution of such, potentially increases the localization accuracy of a query image by image retrieval.

*Visual Localization* is the task of determining the camera pose of one or multiple query images in a specific scene. Visual localization carried out using Convolutional Neural Networks improved in terms of accuracy over the last few years. In general CNNs are trained on training sets containing images of an environment and their corresponding poses. A neural network optimizes its weights by minimizing a loss function. For visual localization this loss function is often based on minimizing pose differences or reprojection errors. Visual localization may benefit by expanding the training sets with a more variable and higher distribution of images and poses. Again, we extend these training sets by utilizing a 3D model to render new images and apply image translation to transform rendered images into a more realistic *captured* domain.

In this contribution, (i) we render images with novel poses from 3D models to increase the quantity of training images. These training images show artifacts and are of limited use for further image analysis. Therefore, (ii) we improve the quality of the training images by image translation. Furthermore, (iii) we show that image-to-image translation concerning 3D models enhances performance of common computer vision tasks.

- Feature matching is significantly increased by translated images compared to rendered images.
- Image retrieval concerning translated images provides clearly better results in contrast to captured images.
- Visual localization is improved by augmenting captured images with translated images. Furthermore, training only on translated images performs comparable to training on captured images.

This contribution is organized as follows. After reviewing related work on image-to-image translation, feature matching, image retrieval and visual localization in Section 2, the utilized methods are depicted in Section 3. The performed experiments on feature matching, image retrieval and visual localization are introduced in Section 4. We discuss the experiments and their outcome in Section 5 and conclude and give an outlook for future research in Section 6.

## 2. RELATED WORK

Approaches to augment training data sets are well established in the field of computer vision (Gharbi et al., 2016; Lemley et al., 2017). Data augmentation boosts performance in classification (Ng et al., 2015), segmentation (Rajpura et al., 2017), object recognition (Maturana & Scherer, 2015), object detection (Peng et al., 2015), hand gesture estimation (Molchanov et al., 2015), camera pose regression (Mueller et al., 2018) or human pose estimation (Rogez & Schmid, 2016). Learning based methods and CNNs can be trained to improve handling invariances like translation or rotation which helps for generalization of the networks (Parkhi et al., 2015). Furthermore augmenting training data by generating synthetic images is known as a valuable process of data augmentation. Synthetic images of text in clutter are generated to train a Fully-Convolutional Regression Network (FCRN) (Gupta et al., 2016). For efficient view registration with respect to a point cloud, synthetic views are generated to enhance the registration of images taken from novel view points (Irschara et al., 2009).

*Image-to-Image Translation* on paired training data has recently been addressed to convert input images from one domain into another, like *gray-scale* to *color* (Iizuka et al., 2016), *day* to *night*, *aerial* to *map* and others (Isola et al., 2017). These translations rely on training sets of aligned image pairs - so-called paired training data. Image translation trained on unpaired data has been addressed for artistic style transfer (Johnson et al., 2016; Gatys et al., 2016) or other domain translations like *horse* to *zebra* or *summer* to *winter* (Zhu et al., 2017). Such

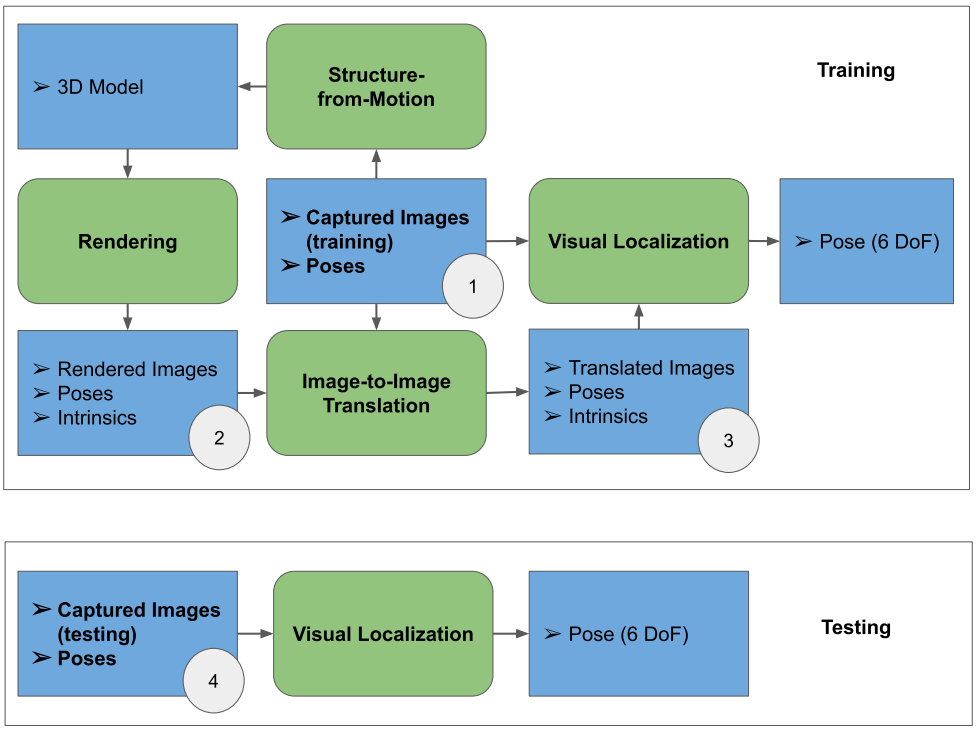


Figure 1. Workflow of translating images from a training set of captured images and training a visual localization pipeline. Captured training images ① are used to create a 3D model through Structure-from-Motion (SfM). Rendered images are generated with specific poses and with camera intrinsics within this 3D model ②. These rendered images and the captured images serve as input to train an image translation network. This network translates images in the *rendered* domain into translated images in the *captured* domain ③. Captured ① and translated ③ images are used to train a visual localization pipeline. Experiments are also carried out on rendered images ② and on feature matching as well as image retrieval. Testing is carried out on the captured images from a test sequence ④. None of the test images is utilized in the prior training process. For comparison purpose each experiment is carried out on the captured images, the rendered images and the translated images in Section 4. Given data is highlighted in bold style.

image translation showed beneficial impact for feature matching and image retrieval translating nighttime to daytime images (Anoosheh et al., 2019; Porav et al., 2018). With recent research the number of domains is extended to numerous, e.g. 16 translations between artistically styles or four domains for translations between the seasonal domains as *spring*, *summer*, *autumn* and *winter* (Anoosheh et al., 2018). These translations are predominantly carried out utilizing Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). Adversarial networks are also used to generate training data by transforming rendered images of eyes to more realistic samples for eye gaze estimation (Shrivastava et al., 2017).

*Feature Matching* is one of the most fundamental algorithms in computer vision. There are several established algorithms in this context like SIFT (Lowe, 2004), SURF (Bay et al., 2006) or ORB (Rublee et al., 2011). Numerous computer vision challenges can be tackled by the support of these algorithms, e.g. image classification (Bosch et al., 2006), object detection (Li & Zhang, 2013), tracking (Zhou et al., 2009), 3D reconstruction (Schönberger & Frahm, 2016), Simultaneous Localization and Mapping (Mur-Artal et al., 2015) or visual localization (Sattler et al., 2017). It is shown that rendering images in point clouds created by laser scans and images improved feature matching and visual localization (Sibbing et al., 2013). However, in contrast to their work our 3D models are reconstructed only by images and are less detailed. Therefore, their techniques are not applicable on our data. Aerial images are matched to terrestrial images using rendered images of a 3D model (Shan et al., 2014). Generating the rendered images from a wide distance compensates the quality of the rendering.

*Image Retrieval* became popular with the emergence of large-scale image collections. Content Based Image Retrieval (CBIR) considers colors, shapes or textures to associate a query image to its most similar image(s) in a training set. There are several approaches available to tackle this task. Solutions utilize grey values (Schmid & Mohr, 1997), Eigenfeatures (Swets & Weng, 1996), VLAD (Jégou et al., 2010) - a compact descriptor to make image retrieval more efficient concerning run time and storage (Arandjelovic & Zisserman, 2013) - or Convolutional Neural Networks (Sharif Razavian et al., 2015; Babenko & Lempitsky, 2015). Image retrieval was improved for situations where the scene appearance changed due to variable illuminations over time by generating virtual views from Google street-view panoramas (Torii et al., 2015). In contrast to our work, only individual depth maps and no global 3D model are used.

*Visual Localization* by pose regression with Convolutional Neural Networks was introduced with the publication of PoseNet (Kendall et al., 2015). Further development of loss functions (Kendall & Cipolla, 2017) or the implication of Long-Short Term Memory (Walch et al., 2017) boosted the performance of image-based localization. Other research focuses on transferring pose regression from large to small networks reducing memory requirements (Mueller et al., 2017). Data augmentation is tackled by adding rendered images to the training data to improve performance of a pose regression pipeline (Mueller & Jutzi, 2018). The first work on scene coordinate regression for camera relocalization is based on random forests rather than deep learning (Shotton et al., 2013). Latest developments are combining deep learning and the well-known

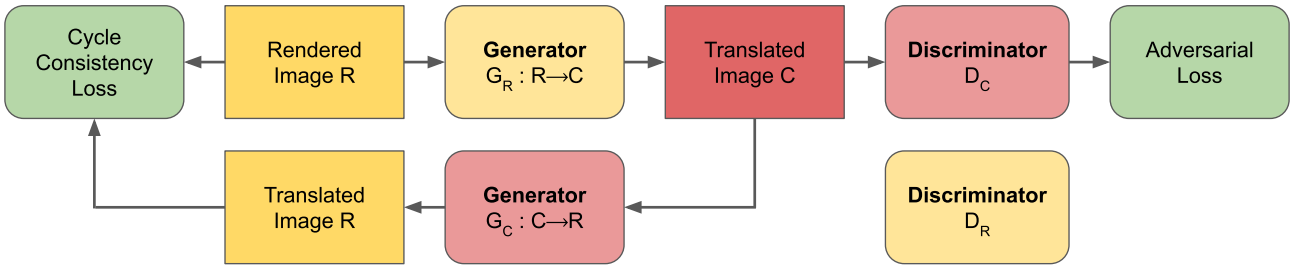


Figure 2. Overview of the utilized GAN system. The training pass for the direction from a rendered image to a captured image $R \rightarrow C$ is shown. Besides the Adversarial Loss, a Cycle Loss is utilized to encourage inverse mappings such that $G_R(G_C(c)) \approx c$. The training pass for the opposite direction $C \rightarrow R$ is executed likewise. Discriminator $D_R$ is illustrated for completeness.

perspective-n-point problem (Haralick et al., 1994) to regress 6-Degree of Freedom (DoF) poses from images (Brachmann & Rother, 2018). The pipeline firstly regresses scene coordinates by a CNN and subsequently applies DSAC (Brachmann et al., 2017), a framework of differentiable RANSAC (Fischler & Bolles, 1981), for finding 2D-3D matches followed by a pose hypothesis estimation. This hybrid approach scores better results as hand-crafted approaches on visual localization. The interest of visual localization in challenging environments with changing weather, daylight or seasonal conditions is important for the navigation of self-driving vehicles and the localization for augmented-reality applications. Therefore, data sets covering these characteristics were published recently (Sattler et al., 2018). Paintings and historical photographs are matched to a 3D model for pose estimation, whereas features are learned to match between paintings and rendered images (Aubry et al., 2014).

Rather than learning to extract similar features, we aim to adjust rendered images to fit our target domain. Investigation on CNN-based pose regression showed that no current pose regression approach outperforms handcrafted retrieval methods consistently (Sattler et al., 2019). We aim to enhance such visual localization approaches by data augmentation with image translation.

## 3. METHODOLOGY

In the methodology section we focus on Image Translation (Section 3.1) and common computer vision tasks, like Feature Matching (Section 3.2.1), Image Retrieval (Section 3.2.2) and Visual Localization (Section 3.2.3).

The general workflow of translating images from a training set of captured images and employing them to the selected tasks is shown in Figure 1 on the example of visual localization. A training set of captured images (① in Figure 1) is used to create a 3D model through Multi-View Stereo (Schönberger et al., 2016). The model is used to render images (Waechter et al., 2017) with specific poses and camera intrinsics ②. These rendered images and the captured images serve as input for training an image translation network. The trained network then translates images from the *rendered* domain to the *captured* domain ③. Captured ① and translated ③ images are used to train a visual localization pipeline. Experiments are also carried out on rendered images ② and on feature matching as well as image retrieval. Testing is carried out on the captured images from a separate test sequence ④. None of the test images is utilized in the prior training process. For comparison purpose, each experiment is carried out on the captured images, the rendered images and the translated images in Section 4. The *Shop Façade* data set from the *Cambridge Landmarks* benchmark (Kendall et al., 2015) serves for these experiments. This data set has a spatial extension of approximately $25m$ x $35m$. The scene mainly shows the façade of a shop. The training set consists of 231 images and their corresponding poses, whereas the test set consist of 103 images and poses.

### 3.1 Image Translation

Image translation is carried out by utilizing ToDayGAN (Anoosheh et al., 2019), a Generative Adversarial Network (GAN) based on CycleGAN (Zhu et al., 2017). GANs generally consist of two independent neural networks which compete with each other. A so-called generative network generates synthetic images while a discriminative network tries to distinguish between real images and the synthetic data, that is the output of the generator network. This procedure allows to generate a vast amount of synthetic data while retaining a realistic appearance and thus serves for data augmentation. The image translation networks perform a mapping of images between two domains $C$ and $R$, corresponding to the *captured* and *rendered* domain. Unpaired samples of both domains $c_i$ and $r_j$, where $i = 1...N$ and $j = 1...M$ are provided during training. An alignment of training samples is not necessary due to the cycle consistency loss introduced in CycleGAN. The network consists of two generators $G_R : R \rightarrow C$ and $G_C : C \rightarrow R$ to translate images between the domains as well as two discriminators $D_R$ and $D_C$ to distinguish between translated and captured images. The GAN is trained for minimizing both, an adversarial loss and a cycle consistency loss (Figure 2). The cycle consistency loss specifies the constraint in such a way that a translation $R \rightarrow C$ followed by $C \rightarrow R$ is hold to lead to the same image as the original input image.

$$G_R(G_C(c)) \approx c \tag{1}$$

For our purpose on augmenting the *Shop Façade* data set, we translate rendered images from the *rendered* domain to the *captured* domain. Therefore, the rendered-to-captured generator $G$ is used. The training images from the *Shop Façade* data set serve as training samples for the *captured* domain. Images rendered from multiple poses in the 3D model of the scene (Sattler et al., 2019) serve as training samples for the *rendered* domain. The 3D model is generated by COLMAP's SfM pipeline (Schönberger & Frahm, 2016; Schönberger et al., 2016). Poses for rendering additional images are generated in a grid with a spacing of 25 cm. Poses are only generated up to 3 meters away from the nearest original training pose. The orientation of each new pose is set to the orientation of the nearest training image. Thereby, additional poses have been generated to render images from new positions and with different points of

view. In total 2652 rendered images and the 231 captured images build the training data for training the image translation network. Figure 3 shows synthetic generated poses of the rendered images (dark red), training poses of captured images (red) and test poses of the captured images (green). In this context Figure 4 shows an image rendered from the 3D model. Figure 5 shows the same image translated into the *captured* domain by the image translation network. The rendered images as well as the translated images are available online[1].

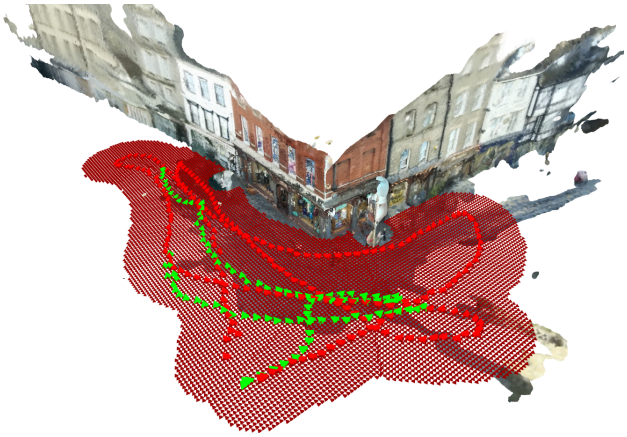

Figure 3. Visualization of synthetic generated poses for the rendered respectively translated images (dark red), poses of of the captured training images (red) and poses of the captured test images (green).

Figure 4. Example of a rendered image from the 3D model. The 3D model is generated from the captured training images.

Figure 5. Example of a translated image from the *rendered* domain into the *captured* domain by image translation. This image was translated from the rendered image shown in Figure 4.

### 3.2 Computer Vision Tasks

We evaluate the impact of using image translation on different computer vision tasks, namely feature matching, image retrieval and visual localization.

[1] https://github.com/tsattler/understanding_apr

**3.2.1 Feature Matching** As one of the most important and fundamental problems in image processing, we perform feature matching for evaluating the quality of image translation. We measure the performance of feature matching based on the number of inliers between images from a training data set and images from the test data set. We depict the inliers within a geometric similarity transformation (Hartley & Zisserman, 2003) and use a variant of MLESAC (Torr & Zisserman, 2000) for model fitting. Feature detection and description is implemented using Speeded-Up Robust Features (SURF) (Bay et al., 2006). Feature matching is then performed by an approximate nearest neighbour search (Muja & Lowe, 2009). To ensure matching by an overlapping field-of-view between test images and training images we employ Bag of Visual Words (Csurka et al., 2004). Concerning feature matching, only ten nearest neighbours in the training set are considered. As a measure of quality, we take the number of inliers between test images and their nearest training images into account, whereas a higher number of inliers corresponds to a higher matching quality. We perform feature matching on the three training data sets of the *captured*, *rendered* and *translated* domain in the experiments (Section 4.1).

**3.2.2 Image Retrieval** For further evaluation of the feasibility of image translation, we apply image retrieval by using a Bag of Visual Words approach. The goal is to compare single test images to a set of training images and to find the images with the highest similarity. Subsequently, a pose difference is computed by taking the pose of the test image and the poses of the most similar training images into account. An unweighted average of poses is computed if multiple training images are taken into account for pose estimation. Therefore, a visual vocabulary with 250 visual words is created by utilizing SURF to extract features and their descriptors from all training images. All features are clustered by using k-means with 250 clusters, whereby every cluster represents a visual word. Investigations using more visual words did not significantly change the results. Based on these visual words a histogram for every training image is derived. Subsequently the features and descriptors of the test images are derived with the same strategy and added to one of the 250 clusters by using a simple nearest neighbour approach. Adjacent, a histogram of visual words of the test image is derived and compared to the Bag of Visual Words by using histogram intersection. Therewith, the best matching histograms of the images from the training set are identified and assigned to a test image. The images corresponding to these histograms are considered as the nearest neighbours for the test image. To evaluate image translation, in Section 4.2 we investigate the performance of image retrieval on the three different data sets mentioned above, namely captured images, rendered images and translated images. Besides a visual comparison, a geometrical evaluation is carried out as mentioned by computing pose differences between ground truth and estimated poses. The ground truth poses for training and test sets are given from the benchmark data set. The estimated poses again are derived by determining the mean poses of the nearest training images. In detail, the euclidean distance between two poses defines the difference of translation $d$ as

$$d = \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2 \tag{2}$$

where $\mathbf{x}_i$ is the position of a test image and $\hat{\mathbf{x}}_i$ is the position of a training image and $\|\cdot\|_2$ the Euclidean Norm.

The difference of rotation between a test image and a training image $\theta$ is computed by

$$\theta = 2 * arccos\left(\mathbf{q}_i \cdot \frac{\hat{\mathbf{q}}_i}{\|\hat{\mathbf{q}}_i\|_2}\right) \quad (3)$$

where $\mathbf{q}_i$ is the normalized quaternion of a test image and $\hat{\mathbf{q}}_i$ is the quaternion of a training image. $\theta$ therefore depicts the angle between the orientation of a training and a test image.

**3.2.3 Visual Localization** For further investigations on the feasibility of translated images to enhance image analysis we perform visual localization utilizing DSAC++ (Brachmann & Rother, 2018). This approach consists of a neural network and a pose estimation pipeline based on 2D-3D correspondences. The network takes RGB images, their corresponding poses and intrinsic camera calibrations as input for the training procedure and regresses a 6-DoF pose for single test images. Initially a CNN predicts a depth value for every pixel in the input image. This leads to a 2D-3D correspondence from every pixel to a point in the 3D scene. By solving the perspective-n-point problem a camera pose can be estimated. Multiple camera pose hypotheses are computed – each from four of such 2D-3D correspondences. This is followed by a pose hypothesis selection and a pose hypothesis refinement leading to a final pose estimate. The network is optimized by minimizing a pose loss in an end-to-end training using standard backward propagation. Our training sets consist of the captured images from the *Shop Façade* data set, the rendered images generated from a 3D model and the translated images generated by image translation.

## 4. EXPERIMENTS

For evaluating the enhancement of common computer vision tasks with translated images, we investigate Feature Matching (Section 4.1), Image Retrieval (Section 4.2) and Visual Localization (Section 4.3). All experiments are carried out on the *Shop Façade* data set from the *Cambridge Landmarks* (Kendall et al., 2015) visual localization benchmark.

### 4.1 Feature Matching

We investigate the improvement of feature matching on translated images in contrast to feature matching on rendered images. Therefore we extract SURF features from all training images and test images. Since matching every test image to every training image would include matching images without a joint view of the scene, we pre-select the matching candidates by the image retrieval algorithm mentioned in Section 3.2.2 and match every test image to its ten nearest neighbours. Figure 6 depicts results of a test image (left column) matched to a training image from the captured (top), rendered (mid) and translated (bottom) data set (right column). Table 1 shows the average number of matches respectively inliers between the test images and the images of the training sets. Image translation on rendered images increases the number of inliers significantly.

| Data set | Avg. # of matches | Avg. # of inliers | % |
|---|---|---|---|
| Captured | 678 | 244 | 35.9 |
| Rendered | 240 | 12 | 5.0 |
| Translated | 396 | 79 | 19.9 |

Table 1. Average numbers of total matches and inliers between test images and training images. The test images are the captured images from the *Shop Façade* test sequence. The last column shows the percentage of average inliers in relation to the average matches.

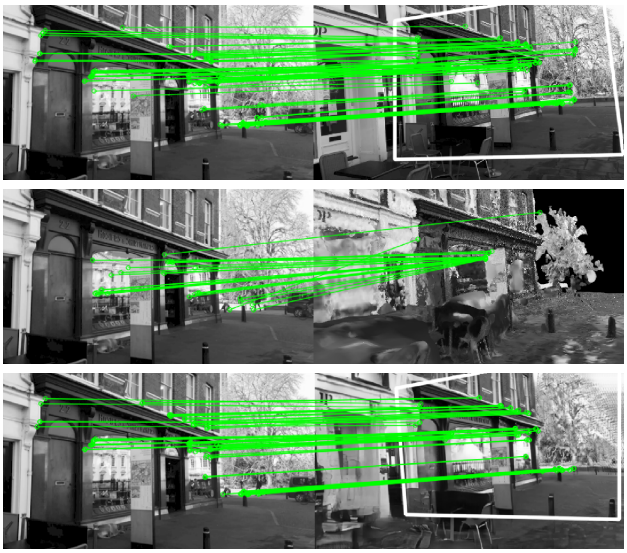

Figure 6. Visual example results for feature matching. Each row shows matched features between a captured image (left) and a training image (right). The white bounding boxes depict the boarders of the projected test images. Training images are from top to bottom in the *captured*, *rendered* and *translated* domain.

### 4.2 Image retrieval

Image retrieval is processed on the captured training images, the rendered images and the translated images. The captured test images serve for evaluation. Figure 7a shows one of the test images, whereas Figure 7b, 7c and 7d each shows the four nearest neighbours of the training sets (captured, rendered, translated) corresponding to the test image. The mean pose differences (unweighted average) are computed between each test image and its top 1, top 4 and top 10 nearest neighbours for each training data set (Table 2). Utilizing the translated images for training clearly leads to better results than utilizing the captured images, potentially benefiting from a denser distribution of training images. Image retrieval on the rendered images performed clearly worse due to high dissimilarity to the test images.

| | Mean Pose Difference | | |
|---|---|---|---|
| Data set | Top 1 | Top 4 | Top 10 |
| Captured | $0.72m/0.43^\circ$ | $0.69m/0.42^\circ$ | $0.82m/0.50^\circ$ |
| Rendered | $2.62m/0.84^\circ$ | $2.73m/0.85^\circ$ | $2.88m/0.89^\circ$ |
| Translated | $\mathbf{0.49m/0.29^\circ}$ | $\mathbf{0.38m/0.28^\circ}$ | $\mathbf{0.49m/0.31^\circ}$ |

Table 2. Mean pose differences of test image poses to their nearest neighbours from the training data sets. The mean pose differences are computed between each test image and its top 1, top 4 and top 10 nearest neighbours for each training data set. Best results are highlighted in bold style.

### 4.3 Visual Localization

For evaluating image translation on visual localization the localization approach presented in Section 3.2.3 is adapted. The data sets of the captured, rendered and translated images serve for training the network. Therefore, the pipeline is trained on each of the mentioned training sets separately. Additionally, a training on a combined training set containing the captured images and the translated images is carried out. All experiments are processed with the same settings, e.g. number of iterations per training step. Testing the networks is carried out on the test set with the captured test images on all four trained

Figure 7. Example results on image retrieval. (a) Shows a test image in the captured domain. (b), (c) and (d) show the 4 nearest neighbours to (a) in the *captured*, *rendered* and *translated* domain.

models. The test results are depicted in Table 3. The network achieved a pose accuracy as median translation and rotation errors of $0.14m/0.7^\circ$ on the captured data, $8.86m/39.5^\circ$ on the rendered data and $0.16m/0.6^\circ$ on the translated data. Training on the combined set of captured and translated images scored $0.12m$ and $0.4^\circ$, which is also the best result.

| Data set | Total # of images | Translation/ Rotation error |
|---|---|---|
| Captured | 231 | $0.14m/0.7^\circ$ |
| Rendered | 2652 | $8.86m/39.5^\circ$ |
| Translated | 2652 | $0.16m/0.6^\circ$ |
| Captured + Translated | 231 + 2652 | $\mathbf{0.12m/0.4^\circ}$ |

Table 3. Median translation and rotation test errors on the captured, rendered and translated data sets. We also trained a model on a combined data set of captured and translated images scoring the best results.

## 5. DISCUSSION

With the experiments on translated images, we show enhancements over the usage of rendered images on feature matching, image retrieval and visual localization. Compared to captured images, the experiments also show promising results on image retrieval and visual localization.

The average number of 12 inliers found on the rendered images is not satisfying for most computer vision task. However, after image translation the average number of inliers increased to 79, which is a decent amount of matches to successfully, e.g. register two images.

Utilizing translated images clearly leads to better results over the usage of the captured images. Image retrieval benefits from the higher number of images leading to a denser sampling compared to captured images, hence finding nearer images and improving the pose estimate. Image retrieval trained on rendered images shows a decreased accuracy compared to retrieval on captured images due to high dissimilarity to the test images.

Image translation for visual localization showed a beneficial impact compared to training on captured data. The network scored similar results as on training with captured data and best results when training on a combined data set of captured and translated images. The network trained on rendered data failed when testing on captured data. That implies that the network potentially learns representations for rendered images, which can not be transferred to captured images. Moreover, we show that image translation can transform these images into valuable training data.

## 6. CONCLUSION AND OUTLOOK

We want to highlight the potential of image translation from a *rendered* domain to a *captured* domain for image retrieval and visual localization. We were able to train a network for visual localization merely on synthetic data (translated images) and achieve similar results compared to training on manually captured data. The accuracy of visual localization improves by training supported with translated images. We additionally mention, that the images of the utilized data set show similar scene views. Bigger gains are possible when translating rendered images from views that are substantially different from the captured views. However, generating plausible translations

for such views is harder, creating the necessity for further research to handle large pose changes between captured and rendered images. Further work on GANs is therefore needed to overcome this issue.

## REFERENCES

Anoosheh A., Agustsson E., Timofte R., Van Gool L., 2018. Combogan: Unrestrained Scalability for Image Domain Translation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 783–790.

Anoosheh A., Sattler T., Timofte R., Pollefeys M., Van Gool L., 2019. Night-to-Day Image Translation for Retrieval-based Localization. *2019 IEEE International Conference on Robotics and Automation*.

Arandjelovic R., Zisserman A., 2013. All about VLAD. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1578–1585.

Aubry M., Russell B. C., Sivic J., 2014. Painting-to-3D Model Alignment via Discriminative Visual Elements. *ACM Transactions on Graphics*, 33, 14.

Babenko A., Lempitsky V., 2015. Aggregating Local Deep Features for Image Retrieval. *Proceedings of the IEEE International Conference on Computer Vision*, 1269–1277.

Bay H., Tuytelaars T., Van Gool L., 2006. SURF: Speeded-Up Robust Features. *European Conference on Computer Vision*, 404–417.

Bosch A., Zisserman A., Muñoz X., 2006. Scene Classification via pLSA. *European Conference on Computer Vision*, Springer, 517–530.

Brachmann E., Krull A., Nowozin S., Shotton J., Michel F., Gumhold S., Rother C., 2017. DSAC-Differentiable RANSAC for Camera Localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6684–6692.

Brachmann E., Rother C., 2018. Learning Less is More – 6D Camera Localization via 3D Surface Regression. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4654–4662.

Csurka G., Dance C., Fan L., Willamowski J., Bray C., 2004. Visual Categorization with Bags of Keypoints. *Workshop on Statistical Learning in Computer Vision*, 1, No. 1-22, Prague, 1–2.

Fischler M. A., Bolles R. C., 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 24, 381–395.

Gatys L. A., Ecker A. S., Bethge M., 2016. Image Style Transfer Using Convolutional Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2414–2423.

Gharbi M., Chaurasia G., Paris S., Durand F., 2016. Deep Joint Demosaicking and Denoising. *ACM Transactions on Graphics*, 35, 191.

Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y., 2014. Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, 2672–2680.

Gupta A., Vedaldi A., Zisserman A., 2016. Synthetic Data for Text Localisation in Natural Images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2315–2324.

Haralick B. M., Lee C.-N., Ottenberg K., Nölle M., 1994. Review and Analysis of Solutions of the Three Point Perspective Pose Estimation Problem. *International Journal of Computer Vision*, 13, 331–356.

Hartley R., Zisserman A., 2003. *Multiple View Geometry in Computer Vision*. Cambridge university press.

Iizuka S., Simo-Serra E., Ishikawa H., 2016. Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification. *ACM Transactions on Graphics*, 35, 110:1–110:11.

Irschara A., Zach C., Frahm J.-M., Bischof H., 2009. From structure-from-motion point clouds to fast location recognition. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2599–2606.

Isola P., Zhu J.-Y., Zhou T., Efros A. A., 2017. Image-to-Image Translation with Conditional Adversarial Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1125–1134.

Jégou H., Douze M., Schmid C., Pérez P., 2010. Aggregating Local Descriptors into a Compact Image Representation. *IEEE Conference on Computer Vision and Pattern Recognition*, 3304–3311.

Johnson J., Alahi A., Fei-Fei L., 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. *European Conference on Computer Vision*, 694–711.

Kendall A., Cipolla R., 2017. Geometric Loss Functions for Camera Pose Regression with Deep Learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5974–5983.

Kendall A., Grimes M., Cipolla R., 2015. Posenet: A Convolutional Network for Real-Time 6-DoF Camera Relocalization. *Proceedings of the IEEE International Conference on Computer Vision*, 2938–2946.

Lemley J., Bazrafkan S., Corcoran P., 2017. Smart Augmentation Learning an Optimal Data Augmentation Strategy. *IEEE Access*, 5, 5858–5869.

Li J., Zhang Y., 2013. Learning SURF Cascade for Fast and Accurate Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3468–3475.

Lowe D. G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60, 91–110.

Maturana D., Scherer S., 2015. Voxnet: A 3D Convolutional Neural Network for Real-Time Object Recognition. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 922–928.

Molchanov P., Gupta S., Kim K., Kautz J., 2015. Hand Gesture Recognition with 3D Convolutional Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1–7.

Mueller M. S., Jutzi B., 2018. UAS Navigation with SqueezePoseNetAccuracy Boosting for Pose Regression by Data Augmentation. *Drones*, 2, 7.

Mueller M. S., Metzger A., Jutzi B., 2018. CNN-Based Initial Localization Improved by Data Augmentation. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4, 117–124.

Mueller M. S., Urban S., Jutzi B., 2017. Squeezeposenet: Image Based Pose Regression with Small Convolutional Neural Networks for Real Time UAS Navigation. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4, 49-57.

Muja M., Lowe D. G., 2009. Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. *International Conference on Computer Vision Theory and Application*, INSTICC Press, 331–340.

Mur-Artal R., Montiel J. M. M., Tardos J. D., 2015. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, 31, 1147–1163.

Ng J. Y.-H., Hausknecht M., Vijayanarasimhan S., Vinyals O., Monga R., Toderici G., 2015. Beyond Short Snippets: Deep Networks for Video Classification. *IEEE Conference on Computer Vision and Pattern Recognition*, 4694–4702.

Parkhi O. M., Vedaldi A., Zisserman A., 2015. Deep Face Recognition. *BMVC*, 1, No. 3, 6.

Peng X., Sun B., Ali K., Saenko K., 2015. Learning Deep Object Detectors from 3D Models. *IEEE International Conference on Computer Vision*, 1278–1286.

Porav H., Maddern W., Newman P., 2018. Adversarial Training for Adverse Conditions: Robust Metric Localisation Using Appearance Transfer. *IEEE International Conference on Robotics and Automation*, 1011–1018.

Rajpura P., Goyal M., Hegde R., Bojinov H., 2017. Dataset Augmentation with Synthetic Images Improves Semantic Segmentation. *arXiv preprint arXiv:1709.00849*.

Rogez G., Schmid C., 2016. MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. *Advances in Neural Information Processing Systems 29*, 3108–3116.

Rublee E., Rabaud V., Konolige K., Bradski G., 2011. ORB: An Efficient Alternative to SIFT or SURF. *IEEE International Conference on Computer Vision*, 1, 2564–2571.

Sattler T., Leibe B., Kobbelt L., 2017. Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39, 1744–1756.

Sattler T., Maddern W., Toft C., Torii A., Hammarstrand L., Stenborg E., Safari D., Okutomi M., Pollefeys M., Sivic J., Kahl F., Pajdla T., 2018. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sattler T., Zhou Q., Pollefeys M., Leal-Taixe L., 2019. Understanding the Limitations of CNN-Based Absolute Camera Pose Regression. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Schmid C., Mohr R., 1997. Local Grayvalue Invariants for Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 530-535.

Schönberger J. L., Frahm J.-M., 2016. Structure-from-Motion Revisited. *Conference on Computer Vision and Pattern Recognition*.

Schönberger J. L., Zheng E., Frahm J.-M., Pollefeys M., 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. *European Conference on Computer Vision*, Springer, 501–518.

Shan Q., Wu C., Curless B., Furukawa Y., Hernandez C., Seitz S. M., 2014. Accurate Geo-Registration by Ground-to-Aerial Image Matching. *2nd International Conference on 3D Vision*, 1, 525–532.

Sharif Razavian A., Sullivan J., Maki A., Carlsson S., 2015. A Baseline for Visual Instance Retrieval with deep Convolutional Networks. *International Conference on Learning Representations*.

Shotton J., Glocker B., Zach C., Izadi S., Criminisi A., Fitzgibbon A., 2013. Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shrivastava A., Pfister T., Tuzel O., Susskind J., Wang W., Webb R., 2017. Learning From Simulated and Unsupervised Images Through Adversarial Training. *IEEE Conference on Computer Vision and Pattern Recognition*.

Sibbing D., Sattler T., Leibe B., Kobbelt L., 2013. Sift-realistic rendering. *2013 International Conference on 3D Vision*, 56–63.

Swets D. L., Weng J. J., 1996. Using Discriminant Eigenfeatures for Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18, 831-836.

Torii A., Arandjelovic R., Sivic J., Okutomi M., Pajdla T., 2015. 24/7 Place Recognition by View Synthesis. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1808–1817.

Torr P. H., Zisserman A., 2000. MLESAC: A new Robust Estimator with Application to Estimating Image Geometry. *Computer Vision and Image Understanding*, 78, 138–156.

Waechter M., Beljan M., Fuhrmann S., Moehrle N., Kopf J., Goesele M., 2017. Virtual Rephotography: Novel View Prediction Error for 3D Reconstruction. *ACM Transactions on Graphics*, 36, 45a.

Walch F., Hazirbas C., Leal-Taixe L., Sattler T., Hilsenbeck S., Cremers D., 2017. Image-Based Localization using LSTMs for Structured Feature Correlation. *Proceedings of the IEEE International Conference on Computer Vision*, 627–637.

Zhou H., Yuan Y., Shi C., 2009. Object Tracking using SIFT Features and Mean Shift. *Computer Vision and Image Understanding*, 113, 345–352.

Zhu J.-Y., Park T., Isola P., Efros A. A., 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *IEEE International Conference on Computer Vision*.