From Dusk till Dawn: Modeling in the Dark

Filip Radenović[1] Johannes L. Schönberger[2,3] Dinghuang Ji[2]
Jan-Michael Frahm[2] Ondřej Chum[1] Jiří Matas[1]

[1]CMP, Faculty of Electrical Engineering, Czech Technical University in Prague
[2]Department of Computer Science, The University of North Carolina at Chapel Hill
[3]Department of Computer Science, Eidgenössisch Technische Hochschule Zürich

Abstract

Internet photo collections naturally contain a large variety of illumination conditions, with the largest difference between day and night images. Current modeling techniques do not embrace the broad illumination range often leading to reconstruction failure or severe artifacts. We present an algorithm that leverages the appearance variety to obtain more complete and accurate scene geometry along with consistent multi-illumination appearance information. The proposed method relies on automatic scene appearance grouping, which is used to obtain separate dense 3D models. Subsequent model fusion combines the separate models into a complete and accurate reconstruction of the scene. In addition, we propose a method to derive the appearance information for the model under the different illumination conditions, even for scene parts that are not observed under one illumination condition. To achieve this, we develop a cross-illumination color transfer technique. We evaluate our method on a large variety of landmarks from across Europe reconstructed from a database of 7.4M images.

1. Introduction

Image retrieval and 3D reconstruction have made big strides in the past decade. Recently, image retrieval and Structure-from-Motion (SfM) methods have been combined to achieve modeling from 100 million images [10]. Combining them can not only tackle scale but also allows to reconstruct spatially complete models with high levels of detail [21]. A key observation is that an increasing number of images in the collections ease the registration of images taken under very different illumination conditions into a single 3D model. A feat that is not achieved by direct matching techniques, but rather by discovering sequences of matching images with a gradual change of the illumination, see Figure 2 for such a transition sequence.

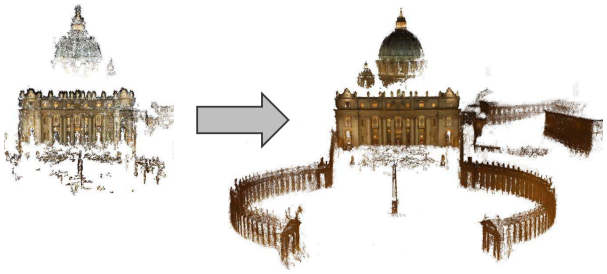

Figure 1. Night model of St. Peter's Cathedral in Rome reconstructed by our method. Left: Model obtained from night images only. Right: Fused, recolored model from day and night images.

A sparse 3D reconstruction of feature points, obtained from a mixed set of day and night images, is reliable and naturally occurs in large-scale photo collections. This is due to the presence of "transition" images and due to the fact, that some of the detected features after photometric normalization provide sufficiently stable matches across illumination transitions. Examples of such feature points, the corresponding image patches, and their normalized descriptor patches are shown in Figure 4.

However, while beneficial for SfM, the registration of mixed illumination images creates challenges for dense 3D reconstruction, which delivers poor results or even fails in the presence of day and night images [15]. In particular, mixed illuminations cause erroneous dense correspondences due to accidental photo-consistency in multi-view stereo that distort the texture composition of the models.

As a *first* contribution of the paper, we propose a method for automatically separating day and night images based on the sparse scene graph produced by SfM and a learned day/night color model. The separated sets of day and night images then allow to compute reliable dense reconstructions for each of the two modalities separately.

This work was done while J. L. Schönberger was at the University of North Carolina at Chapel Hill. We thank Marc Eder for helping with experiments. F. Radenović, O. Chum and J. Matas were supported by the MSMT LL1303 ERC-CZ and GACR P103/12/G084 grants. J. L. Schönberger, D. Ji and J.-M. Frahm were supported in part by the National Science Foundation No. IIS-1349074, No. CNS-1405847, and the MITRE Corp.

Figure 2. Tyn Church, Prague. Registration of day and night images into the same model through smoothly varying illumination in intermediate images during dusk and dawn.

While two separate models on first sight may be seen as a drawback, we demonstrate that they often contain regions in which only one of the models provides reliable surface reconstruction. As expected, we observe that usually daytime images are significantly more frequent and, due to better illumination conditions, lead to overall superior models over nighttime models. Interestingly, we observed several situations where night images provide better reconstruction than their daytime counterparts: (i) when lights at night illuminate or texture areas that are shadowed or ambiguous during the day, and (ii) when areas with repeated and confusing textures are not lit during the night, allowing unambiguous dense matching in those areas. Our *second* contribution is to fuse the initially separated dense models into a superior model combining the strengths of both modalities.

Finally, as a *third* contribution, we introduce a method of color transfer to consistently re-color the composite 3D areas for each illumination condition, even for areas that were not reconstructed under the illumination, *i.e.*, we will compute a nighttime color even for geometry that is only reconstructed in the day model.

In summary, our contributions achieve a more complete and accurate dense 3D reconstruction for mixed day- and nighttime images that are typically present in Internet photo collections. Previously, the joint modeling of day and nighttime images caused disturbing artifacts or even lead to reconstruction failures. Additionally, we are able to reconstruct a complete color representation for the dense model surfaces leveraging the corresponding appearance characteristics of the daytime and nighttime images.

2. Related Work

The seminal paper of Snavely *et al*. [23, 24] first proposed reconstruction from unordered Internet photo collections. To determine overlapping views, Snavely *et al*. performed exhaustive pairwise geometric verification. While this ensures the highest possible discovery rate, it impairs the scalability of their system due to the quadratic complexity growth in the number of images. During the following years, several methods for tackling scalability of unordered photo collection reconstruction were proposed: appearance-based clustering methods for grouping the images [12, 6], vocabulary tree based approaches [1, 14], and most recently streaming based methods leveraging augmented appearance indexing [10]. Although the systems successfully scaled the reconstruction to tens of millions of images, they lost the ability to reconstruct details of the scene in the process. Recently, Schönberger *et al*. [21] proposed a method to overcome this limitation of not being able to reconstruct details. Their method leverages a tightly-coupled SfM and image retrieval system [17] to overcome the loss of fine details in the models while keeping the scalability of the state-of-the-art reconstruction systems. Our reconstruction system is inspired by this method. Snavely *et al*. [23, 24] empirically observed the difficulty in registering night images due to their noisiness and darkness. In our system, we overcome this limitation by registering night images mainly through transition images under intermediate illumination conditions during dusk and dawn (see Figure 2). Snavely's system [22] provided an option to manually select day or night images to explore similar viewpoints and illuminations. In contrast, our system automatically classifies and clusters day and night images. In addition, we use the clustering to improve reconstruction results.

Schindler and Dellaert [19] proposed a method for analyzing the point in time at which a photo was taken. In contrast to our approach, their method was relying on observable changes of the scene geometry, *e.g.*, construction or demolition of buildings, which typically happens over longer periods of time. Our method focuses on modeling the illumination changes over the course of a day. Recently, Matzen *et al*. [16] proposed an approach to model and extract temporal scene appearance changes in 3D reconstructions. They perform temporal segmentation of the 3D model to obtain objects whose appearance changed over time. The recovered object appearance changes (wall art, signs, billboards, storefronts, *etc*.) relate to scene texture changes but not to illumination changes due to their search of change over longer periods of time. In contrast, our algorithm aims at determining periodic short term (over the course of a day) temporal scene appearance and illumination changes. Hence, our proposed approach deals with much smaller appearance differences in segmented parts of the reconstruction. These changes are caused by different illuminations during daytime and nighttime and are not correlated with scene texture changes.

Martin-Brualla *et al*. [15] proposed to compute time-lapse mosaics from unordered Internet photo collections of landmarks. They observed the difficulties posed by the presence of night and day images in the same reconstruction. Specifically, they noted that mixing day and night images within the same model introduces "unrealistic twilight effects". In this paper, we propose an approach that overcomes these failure cases and obtains a correct representation of the 3D model for both modes of illumination.

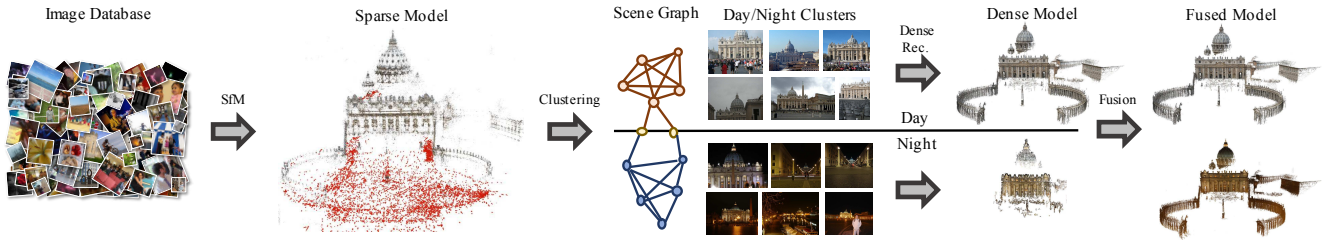

Figure 3. The proposed day/night modeling pipeline starting with sparse modeling to day-night clustering and the final dense modeling.

Ji *et al*. [11] proposed a system to automatically create illumination mosaics for a given outdoor scene from Internet photos. Their work strives to depict temporal variability of the observed scene by presenting a 2D image of the scene with varying illumination along the rows of the image. They perform a search for a chain of images that exercise smooth illumination variation and that are all related through a homography mapping. In contrast, our method considers all available images and not only the images related through homographies. Instead of illumination modeling in 2D, our approach achieves illumination separation and modeling in 3D for the entire scene. Moreover, the ordering of Ji *et al*. [11] heavily relies on the color of the sky shown in the images. Whereas our system can perform day-night separation even with no sky present in any of the images.

Veride *et al*. [25] learned a feature detector which is stable under significant illumination changes, facilitating the matching between day- and nighttime images. They observed that standard feature detectors exhibit significant temporal sensitivity, *i.e*., reduced repeatability under different illumination conditions. We exploit this temporal sensitivity "flaw" of the standard detectors to efficiently split a given 3D model into groups of cameras and points that have the highest illumination change across groups, *i.e*., a group for the day and another for the night.

3. Overview

Before delving into the details of our method for day and night model reconstruction, we provide an overview as illustrated in Figure 3. It starts with a database of unordered images. During the initial phase of the reconstruction, we build a sparse 3D model using SfM (see Section 4). In support of sparse modeling, we index all images in the database using a min-Hash and find reconstruction seeds by leveraging geometrically verified hash-collisions. Next, our SfM algorithm uses these seeds to build sparse 3D models for the scenes contained in the photo collection. Specifically, it uses a feedback loop to gradually extend the reconstruction by dedicated queries against the database. The resulting sparse model contains day and night images registered into the same model and represented as one scene graph.

In the next step, a dense scene model is obtained. Given the previously observed difficulties and artifacts caused by mixed day and night images, we deviate from the standard approach of directly proceeding to dense reconstruction. We first split the scene graph into day and night clusters to separate the images of the different illumination conditions (see Section 5). This in essence separates the scene graph into two scene graphs – one for daytime images and one for nighttime images. Then, we perform separate dense geometry estimation for the images in each of the scene graphs yielding two separate dense 3D models (see Section 6). Subsequently, the two dense models are aligned into one common model representing the overall dense scene geometry. As part of computing the dense scene geometry, we obtain the color information of the point cloud under the two illumination conditions, *i.e*., a daytime color and a nighttime color for each point. Given that not all parts of the common model are necessarily visible both at daytime and at nighttime, we then determine the missing color information through cross-illumination transfer. Specifically, we use one illumination condition to find similar patches with a corresponding color in the other illumination. The color information of the patches under one illumination is then used to compose the missing color information for the point under the other illumination.

4. Reconstruction

In this section, we detail our approach for efficiently reconstructing all 3D models contained in a given image database. We use a generic database from [21] with over 7.4 million images downloaded from Flickr through keywords of famous landmarks, cities, countries, and architectural sites. The approach starts with an initial clustering procedure to find putative spatially related images. These spatially related images are subsequently used to seed an iterative reconstruction process that repeatedly extends the 3D model through a tight integration of the image retrieval and SfM module similar to the approach by Schönberger *et al*. [21, 20]. In contrast to their system, our approach exhaustively builds models for the entire image database. Due to the massive number of images in the database, exhaustive reconstruction imposes several challenges in terms of efficiency, which we address through an initial clustering procedure and a parallelized implementation of their system.

Clustering To seed our iterative reconstruction process efficiently, we find independent sets of spatially overlapping images using the clustering approach by Chum *et al*. [4].

This approach first indexes all database images in a min-Hash table and then uses spatially verified hash collisions as cluster seeds. Next, an incremental query expansion [5, 18] with spatial verification extends the initial clusters with additional images of the same landmark. The nearest-neighbor images in this query expansion step then define the graph of overlapping images, the so-called scene graph. Given that query expansion is a depth first search strategy, the resulting scene graph is only sparsely connected. However, in order to achieve a successful reconstruction, SfM requires a denser scene graph than provided by the clustering method. Therefore, we first densify the scene graph as described in the following paragraph before using it in SfM. Compared to the approach in [21], which takes a single query image as input for the reconstruction, this clustering step reduces the number of query images dramatically. Rather than seeding the reconstruction with 7.4M query images, the clustering procedure identifies 19,546 individual landmarks used to initialize the subsequent reconstruction procedure and thereby reduces the number of seeds by 3 orders of magnitude.

Densification Next, we densify the initially sparse scene graph for improved reconstruction robustness and completeness. In the spirit of Schönberger *et al*. [21], we leverage the spatially verified image pairs and their visual word matches along with an affine model to serve as hypotheses for subsequent exhaustive feature matching and epipolar verification. From this exhaustive verification, we not only obtain a higher number of feature correspondences but we also determine additional image pairs to densify the scene graph. More importantly, beyond the benefit of additional image pairs, the significantly increased number of feature correspondences is essential for establishing feature tracks from day to night images through dusk and dawn. Only through these transitive connections, we are able to reliably register day and night images into a single 3D model.

Structure-from-Motion The densified scene graph is the input to the subsequent incremental SfM algorithm, which treats each edge in the graph as a putative image pair for reconstruction and attempts to reconstruct every connected component within a cluster. Connected components with less than 20 registered images are discarded for the purposes of day/night modeling as they typically lack a sufficient number of transition images during dusk and dawn.

Extension To boost registration completeness, a final extension step issues further queries for all registered images in each reconstructed connected component. If new images are found and spatially verified, we again perform scene graph densification and use SfM to register the new views into the previously reconstructed models. While significantly increasing the size of the reconstructed models, the extension process also improves the performance of the day/night modeling step. Typically, the initial set of images obtained in clustering often only contains images from one modality, *i.e*., either day or night, even though our large-scale image database contains images of both modalities for almost all landmarks. The iterative extension overcomes this problem by incrementally growing the model from day to night or vice versa through transition images during dusk and dawn (see Figure 2 for an example).

5. Day/Night Clustering

After the exhaustive 3D reconstruction stage of all landmarks in the database, we proceed with clustering the images inside each of the 3D models into two groups: day- and nighttime. For crowd-sourced data, the clustering cannot simply rely on embedded EXIF time stamps. In our experiments, the majority of images either have no time stamp information at all or the information is clearly corrupt. We speculate that most images are taken on vacation and people do not adjust the time zone in their cameras. For most landmarks with many registered images, day- and nighttime images are registered into the same model as a result of the extension step (see Section 4). It is well known that standard feature (keypoint) detectors [25] suffer under illumination sensitivity, *i.e*., the reliability of keypoint detectors degrades significantly when the images originate from outdoor scenes during different times of the day or generally different illumination conditions. In this case, the detectors commonly produce keypoints at different locations for day and night lighting conditions [25]. This even holds true when the images are taken from the same viewpoint. Our key insight is to exploit this behavior in order to split the images inside a SfM model into two groups. Our clustering is based on the number of commonly observed 3D points for each pair of images with similar viewpoints. This enables us to identify day and night images registered within a model. For efficient grouping, we leverage a bipartite visibility graph [13], as explained in the following sections.

5.1. Min-cut on Bipartite Visibility Graph

A 3D model produced by SfM can be interpreted as a bipartite visibility graph $\boldsymbol{\mathcal{G}} = (\mathcal{I} \cup \mathcal{P}, \mathcal{E})$ [13], where the images $i \in \mathcal{I}$ and the points $p \in \mathcal{P}$ are the vertices of the graph. The edges of the graph are then defined by the visibility relations between cameras and points, *i.e*., if a point p is visible in an image i, then there exists an edge $(i, p) \in \mathcal{E}$. We define the set of points observed by an image i as:

$$\mathcal{P}(i) = \{p \in \mathcal{P} \mid (i, p) \in \mathcal{E}\}. \quad (1)$$

Our day/night clustering separates the vertices of the graph (the cameras and points) into two groups: one corresponding to day cameras and points and the second for the night cameras and points. More formally, we define two

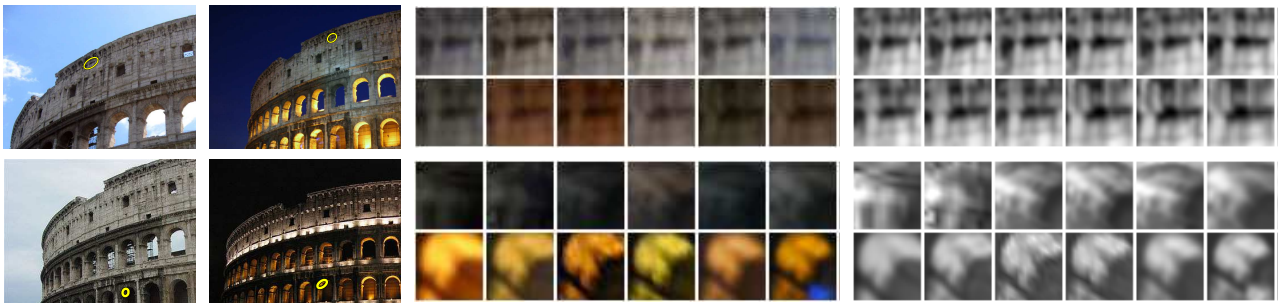

Figure 4. Colosseum, Rome. Two feature tracks containing both day and night images/features. Each row depicts two images labeled as day and night, respectively, followed by a subset of feature patches depicted in two rows, one for day and one for night features, respectively. Intensity normalized patches, grayscale versions used for SIFT description, are shown to the right of the respective color patches. Notice the variation in lighting conditions for day and night, expressed as a significant color difference of patches. Best viewed in color.

label vectors representing the group assignment. Vector $\boldsymbol{\alpha_i}$ for the images and vector $\boldsymbol{\alpha_p}$ for the points:

$$\begin{aligned} \boldsymbol{\alpha_i} &= \{\alpha_i \in \{0,1\} \mid i \in \mathcal{I}\}, \\ \boldsymbol{\alpha_p} &= \{\alpha_p \in \{0,1\} \mid p \in \mathcal{P}\}, \end{aligned} \tag{2}$$

where label variables α_i and α_p correspond to image i and point p, and label $\alpha_i, \alpha_p = 0$ denotes day and label $\alpha_i, \alpha_p = 1$ night. We formulate the problem of separating day from night images as an energy optimization. We propose the following energy function $\mathbf{E}$ over the graph $\boldsymbol{\mathcal{G}}$ that measures the quality of the labeling $\boldsymbol{\alpha_i}, \boldsymbol{\alpha_p}$:

$$\mathbf{E}(\boldsymbol{\alpha_i}, \boldsymbol{\alpha_p}, \boldsymbol{\mathcal{G}}) = \sum_{i \in \mathcal{I}} U_i(\alpha_i) + \sum_{(i,p) \in \mathcal{E}} P_{i,p}(\alpha_i, \alpha_p). \tag{3}$$

The term $P_{i,p}(\alpha_i, \alpha_p)$ describes the pairwise potentials associated with the edges enforcing a smooth labeling of the cameras and points with respect to their mutually observed scene information. A standard Potts model is used for the pairwise potentials, that is $P_{i,p}(\alpha_i, \alpha_p) = 0$ for $\alpha_i = \alpha_p$ and $P_{i,p}(\alpha_i, \alpha_p) = 1$ otherwise. The 3D points incur no unary cost for being assigned either label. The unary cost $U_i(\alpha_i)$ for images is based on the day/night illumination model discussed below. The clustering of all images and points in a model is achieved by minimizing the objective

$$\boldsymbol{\alpha_i}, \boldsymbol{\alpha_p} = \underset{\boldsymbol{\alpha_i}, \boldsymbol{\alpha_p}}{\operatorname{argmin}} \mathbf{E}(\boldsymbol{\alpha_i}, \boldsymbol{\alpha_p}, \boldsymbol{\mathcal{G}}) \tag{4}$$

using the min-cut/max-flow algorithm of Boykov *et al*. [3]. Figure 4 shows examples of 3D point tracks that contain both day and night labels.

5.2. Day/Night Illumination Model

We use a day/night illumination model to estimate the likelihood of an image being taken during day or night respectively. As a feature for the prediction, a spatial color histogram in the opponent color space [9]

$$\begin{aligned} I &= (R+G+B)/3, \\ O_1 &= (R+G-2B)/4 + 0.5, \\ O_2 &= (R-2G+B)/4 + 0.5, \end{aligned} \tag{5}$$

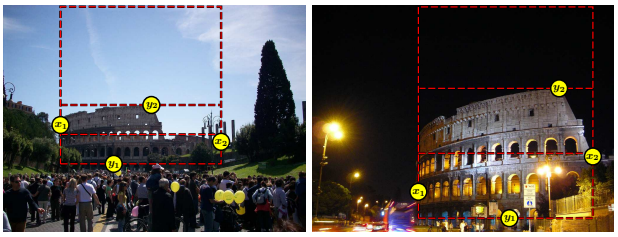

Figure 5. Colosseum, Rome. Examples of image color histogram description area. Coordinates of features reconstructed as 3D points define the bounding boxes used to compute three histograms. Using the 3D model information, we successfully segment out confusing background and are able to focus the description on the three important parts: sky, upper and lower part of the reconstructed landmark.

is used. To reduce the influence of occlusions and background clutter, a three-band spatial histogram is computed over a region of the image directly related to the reconstructed object, as depicted in Figure 5. The bottom two stripes of the histogram equally split the bounding box of feature points that have been reconstructed as 3D points in the model. The top band covers the sky area above the landmark, up to the top edge of the image.

The color is uniformly quantized and each spatial band of the histogram is separately normalized by the number of pixels per region. The final illumination descriptor is obtained by concatenating the color histograms for the three spatial bands. In our experiments, we use $n = 4$ bins per color channel resulting in an image descriptor of dimensionality $D = 3n^3 = 192$.

To classify the illumination descriptors into daytime and nighttime, a linear SVM [2] is trained on ground-truth labeled images of our largest model (Colosseum, Rome). The same trained SVM is used to compute the unary terms for each image i in all reconstructed models:

$$U_i(\alpha_i) = \begin{cases} 0 & \text{if } \alpha_i = \text{SVMp}(i), \\ c \cdot \text{SVMs}(i) \cdot |\mathcal{P}(i)| & \text{otherwise,} \end{cases} \tag{6}$$

where $\text{SVMp}(i)$ and $\text{SVMs}(i)$ denote the SVM's label prediction and the absolute value of the prediction score of im-

age i, respectively. The confidence constant c of the trained SVM has higher confidence for higher values $c > 0$ and in our experiments we set $c = 1$. The cardinality of the set of observed points $\mathcal{P}(i)$ is equal to the number of edges that connect image i to 3D points in the visibility graph.

The label of image i is decided based on the labels of its observed points (pairwise term) and by the confidence of the linear SVM prediction (unary term). In order for this process to be fair for all images, we multiply the SVM score by the number of observed points $|\mathcal{P}(i)|$ for the final unary term. This number defines the percentage of observed points that should have different labels to change the SVM prediction for the image.

6. Day/Night Modeling

After obtaining the image clustering, we first aim to reconstruct the separate models and then combine them into a joint model to produce consistent geometry and texture within each modality, as detailed in this section. Typically, there is an uneven distribution of day and night images, causing one of the modalities to have lower scene coverage. In addition, the different illumination conditions during day and night allow for reconstruction of details that are clearly visible during the day but not at night and vice versa. For example, many landmarks are lit during the night and a reconstruction of fine details is oftentimes possible for night images while during the day those structures are hidden in shadows. Hence, in the second step, we fuse the geometry of the two models in order to obtain better completeness in terms of scene coverage and reconstruction of fine details. To obtain consistent color for the fused model, we recolor the structure of the respective other modality through repainting of visible structure and inpainting of structures not covered by images. The following sections describe our proposed approach in detail.

6.1. Dense Reconstruction

For dense reconstruction, we first separate the sparse model into its day and night modalities based on the labels $\boldsymbol{\alpha_i}$ and $\boldsymbol{\alpha_p}$. For most models, there are enough images during day and night to allow for dense reconstruction in both modalities. We split the graph $\boldsymbol{\mathcal{G}}$ into two disjoint sub-graphs: $\boldsymbol{\mathcal{G}_d}$ for the day modality, and $\boldsymbol{\mathcal{G}_n}$ for the night modality. We separate the tracks of points that are visible in both day and night images. The two graphs serve as the input to the dense reconstruction system by Furukawa and Ponce [7, 8]. Separate reconstruction of day and night images removes many of the disturbing artifacts present when using all images in a model (see Figure 6). To mitigate reconstruction artifacts caused by sky regions, we create segmentation masks using an improved version of the approach proposed by Ji *et al.* [11]. In distinction to their approach, we leverage the sparse point cloud as an additional clue for deciding whether parts of the image belong to the sky or not. The outputs of this step are separate models for day and night. In the next section, we describe an approach that fuses the two models and leverages the benefits of the respective other modality for increased model completeness and detail reconstruction.

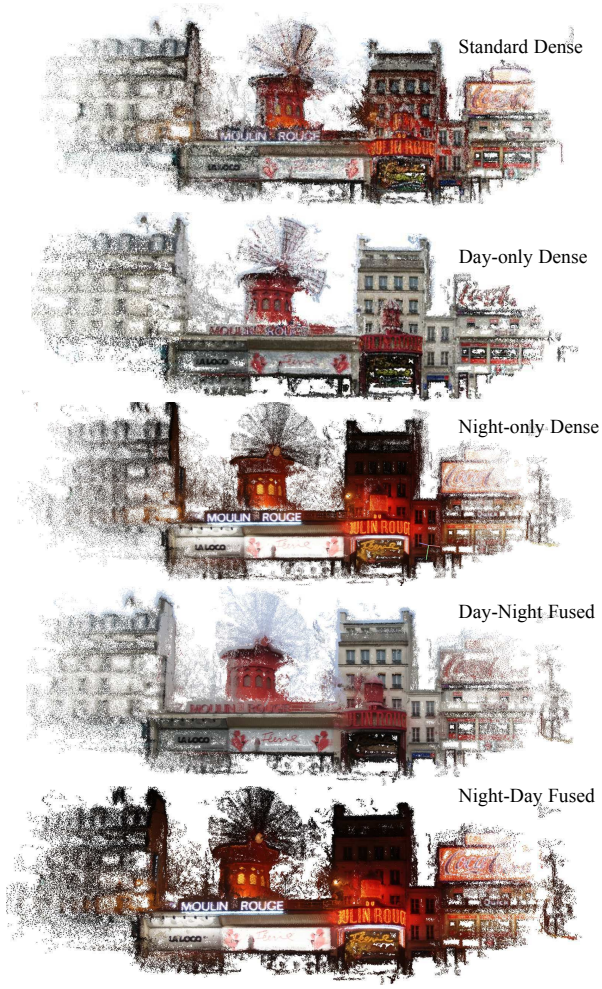

Figure 6. Moulin Rouge, Paris. Standard dense modeling using day and night images creates disturbing artifacts, while a separate modeling for day and night images produces consistent geometry and coloring. Fusion and recoloring improves completeness, appearance, and accuracy.

6.2. Fusion

Typically, the scene coverage of day and night models are very different due to a multitude of reasons. First, parts of the scene may not be covered by any images in one of the modalities, *e.g.*, caused by occlusion or lack of images. In addition, we found that for some scenes, parts of the reconstruction are not covered by images at all during the night due to restricted access in those areas, *e.g.*, the inside of the Colosseum. A second reason for different scene coverage is the different illumination conditions causing dynamic

range issues for the cameras that often prevent reliable reconstruction of scene parts, even though they are theoretically visible. Especially for night images, parts are often under-illuminated or lack any illumination at all. There is a similar issue for day images as well, *e.g.*, shadows caused by intense sunlight often prevent reconstruction of structure. One such case is depicted in Figure 6. Using the default parameters, the dense reconstruction method by Furukawa and Ponce [8] is very conservative in terms of creating dense points, *i.e.*, 3D structure only appears in high confidence areas. Therefore, geometric fusion of the two models enables the use of structure that is more accurately reconstructed from day or night images. As a first step, we perform alignment of the two models into the same reference frame using the correspondences from points that appear both in night and day images. However, such fused models contain both day and night points and thus suffer from inconsistent coloring. In the following paragraphs, we describe a joint repainting and inpainting procedure to color the fused day points in the night model and vice versa (see Figure 7). For simplicity, we explain the procedure for the case of coloring the fused point cloud using the night images, but the approach is analogous in the opposite direction.

Repainting As explained in the previous paragraph, many dense points are reconstructed in day but not in night models, even though they are covered by night images. We project these points into all night images and determine their color as the median of all projections. For occlusion handling, we enforce depth consistency with the sparse point cloud. The depth of the dense points must be within the 10th and 90th percentile of the depth range of the observed sparse points of an image. While this cannot account for fine-grained occlusions, in our experiments, the extracted colors are not affected by occluded observations due to the robust averaging of colors.

Inpainting For those points that are not visible in any night image, we propose a novel inpainting method. The method learns the appearance mapping between known corresponding day and night patches to predict the color of unseen points. To establish dense correspondence between day and night patches, we first project all points into day and night images. Any point that projects both into day and night images defines a correspondence that we use to infer the appearance of a day point during the night. Each of the correspondences usually projects into multiple day and night images. An average color histogram is extracted from a 5×5 patch around the projected image location, for each correspondence between day and night images. While we tried to incorporate shape information as descriptors, we found color histograms to be sufficiently distinctive features and best performing for the task of inpainting. Using these histograms as input, we train a nearest-neighbor regressor to map from day patches to night patches. To inpaint the color

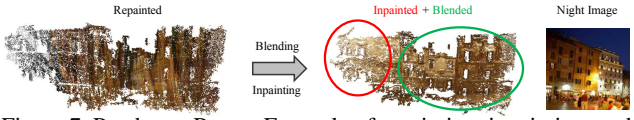

Figure 7. Pantheon, Rome. Example of repainting, inpainting, and blending for building facade that is not present in the original night reconstruction.

of points that only project to day images, we extract the average day color histogram for that point and use our trained regressor to predict its most likely appearance during the night. This inpainting method enables us to obtain a model during the night that is as complete as during the day. In all our experiments, we use $N = 20$ nearest neighbors for the regression and $D = 96$ dimensional histograms for the appearance descriptor.

Blending Even though we are using a robust average in the repainting step, low-coverage points sometimes suffer from abrupt changes in appearance in 3D space whenever the field of view of one image ends. To counteract this artifact, we propose to blend these points by predicting their appearance using the same mapping as in the inpainting step. We improve the color of any point with a track length $t < t_{min}$. The originally repainted color is then blended with the inpainted color based on the track length of the point. The blended color of a point is calculated as

$$\mathbf{c}_{bl} = \frac{t_{min} - t}{t_{min}} \cdot \mathbf{c}_{inp} + \frac{t}{t_{min}} \cdot \mathbf{c}_{rep}, \qquad (7)$$

where $\mathbf{c}_{inp}$ and $\mathbf{c}_{rep}$ denote the inpainted and repainted colors, respectively. In all experiments we set $t_{min} = 10$.

7. Results

After describing our novel approach for day/night modeling, we now evaluate our method on the entire 7.4M image database and present results for a variety of scenes. Our experiments demonstrate that the proposed algorithm robustly generalizes to different illumination conditions.

Reconstruction The iterative reconstruction process for the database of 7.4 million images converges in 3 iterations for all clusters in the database and takes around one week on a single desktop machine. We produce day and nighttime models for any reconstructed cluster that has a sufficient number of registered images, *i.e.*, at least 30 day and 30 night images. We find 1,474 such models out of the initial set of 19,546 clusters used to seed the reconstruction pipeline. These models have 239,717 unique, registered images contained in 845 disjoint landmarks. The average ratio of day to nighttime images in the reconstructions is 9:1.

Clustering To evaluate our clustering approach, we hand-labeled 13,931 images of 6 different landmarks present in the dataset using the two classes of labels "day" and "night" (see Table 1). For the sake of comparison, we also introduce a baseline method for image clustering into day- and nighttime images using k-means clustering with two clusters on

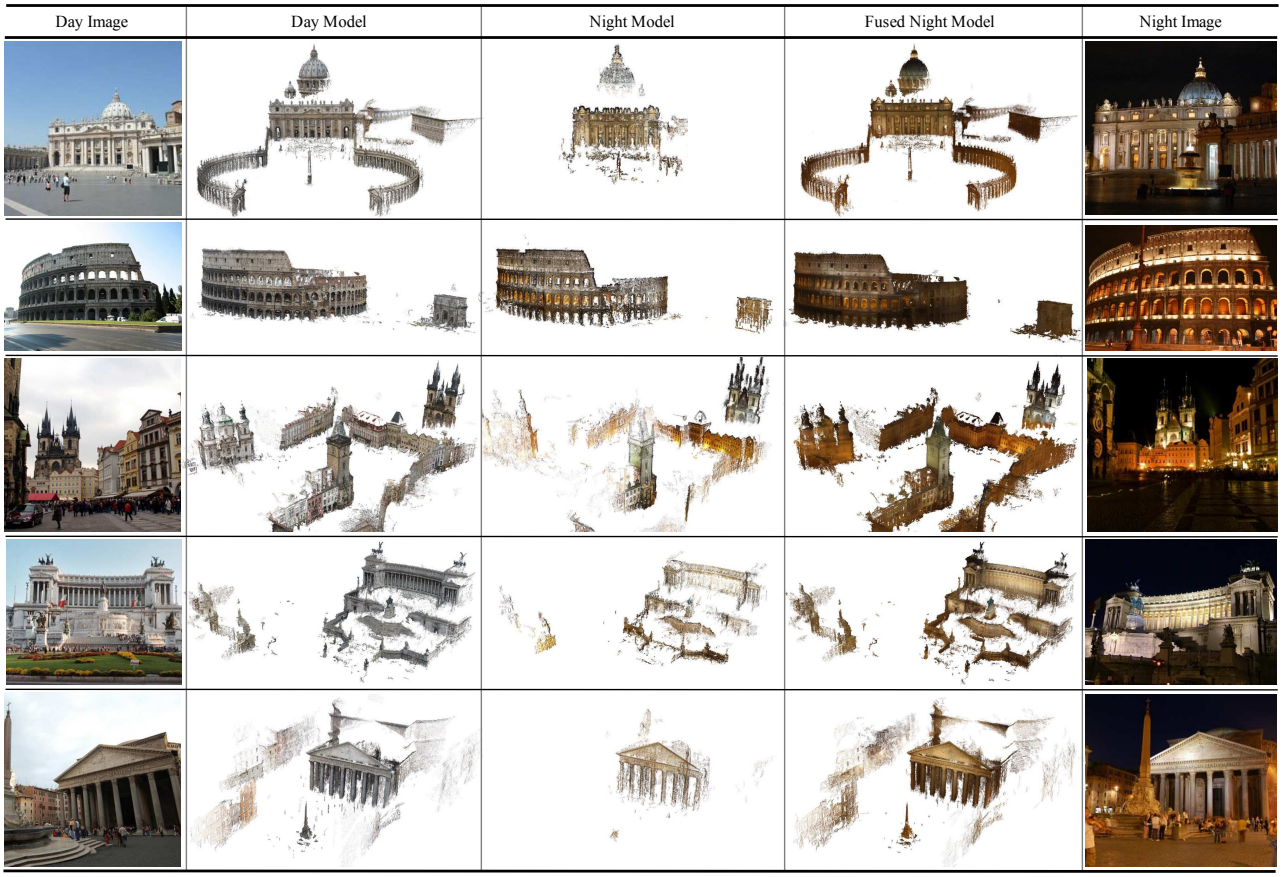

Figure 8. Example of reconstructions produced by our method for St. Peter's Basilica in Vatican, Colosseum in Rome, Astronomical Clock in Prague, Altare della Patria in Rome, and Pantheon in Rome.

the HSV color histograms of the images. Our clustering approach achieves almost perfect classification for the day and night images. Even in the challenging case with only few night images. We outperform k-means on all landmarks and, most importantly, we can classify night images very accurately, which is crucial for avoiding artifacts in day/night modeling. This is even more notable considering that night images are significantly outnumbered in most of the models.

Geometric Fusion Figure 8 impressively demonstrates the improved completeness and accuracy of night models by the geometric fusion. In addition, Figure 6 also depicts an example of the opposite direction, where the structure of day model is improved through the night model. We encourage the readers to view the supplementary material for additional impressions and videos.

Color Fusion Figure 7 demonstrates the proposed repainting, inpainting, and blending method applied to a building facade in a low-coverage part of the Pantheon reconstruction. The structure is not reconstructed in the original night model (Figure 8). Hence, the entire structure consists of repainted points from the day reconstruction. In addition, our method effectively inpaints structure that is not visible in any night images and removes artifacts through blending.

			Ours		Baseline	
Landmark	**# D**	**# N**	**TP**	**FP**	**TP**	**FP**
Spanish Steps	1030	92	98.91	3.26	93.48	14.13
Moulin Rouge	880	754	87.00	0.93	85.81	1.33
Castel St'Angelo	1400	129	99.22	6.20	93.02	6.98
Astronomical Clock	2243	1375	97.89	5.60	80.15	2.98
Altare d. Patria	1993	357	97.76	2.52	92.72	4.20
St. Peter's Basilica	1980	495	98.99	2.22	87.47	6.46

Table 1. Quantitative evaluation of clustering accuracy for night images. Ground-truth labels obtained through manual labeling.

8. Conclusions

We introduced a novel algorithm that handles and benefits from the variety of scene illuminations naturally present in large-scale Internet photo collections. This is in stark contrast to previous methods that treated multiple illuminations as a nuisance or failure condition. We exploit the additional information to obtain a more complete and accurate 3D model and to create multi-illumination appearance information for the 3D model. The proposed method demonstrates that we can leverage the additional information provided by the different illuminations to boost modeling quality for both geometry and appearance.

References

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. Seitz, and R. Szeliski. Building Rome in a Day. *Communications of the ACM*, 2011. 2

[2] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006. 5

[3] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE PAMI*, 2004. 5

[4] O. Chum and J. Matas. Large-scale discovery of spatially related images. *IEEE PAMI*, 2010. 3

[5] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*, 2007. 4

[6] J. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building Rome on a Cloudless Day. In *ECCV*, 2010. 2

[7] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Towards internet-scale multi-view stereo. In *CVPR*, 2010. 6

[8] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE PAMI*, 2010. 6, 7

[9] J.-M. Geusebroek, R. Van den Boomgaard, A. W. Smeulders, and H. Geerts. Color invariance. *IEEE PAMI*, 2001. 5

[10] J. Heinly, J. L. Schönberger, E. Dunn, and J.-M. Frahm. Reconstructing the World* in Six Days *(As Captured by the Yahoo 100 Million Image Dataset). In *CVPR*, 2015. 1, 2

[11] D. Ji, E. Dunn, and J.-M. Frahm. Synthesizing Illumination Mosaics from Internet Photo-Collections. In *ICCV*, 2015. 3, 6

[12] X. Li, C. Wu, C. Zach, S. Lazebnik, and J.-M. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *ECCV*. 2008. 2

[13] Y. Li, N. Snavely, and D. P. Huttenlocher. Location recognition using prioritized feature matching. In *ECCV*. 2010. 4

[14] Y. Lou, N. Snavely, and J. Gehrke. MatchMiner: Efficient Spanning Structure Mining in Large Image Collections. In *ECCV*, 2012. 2

[15] R. Martin-Brualla, D. Gallup, and S. M. Seitz. Time-lapse mining from internet photos. In *SIGGRAPH*, 2015. 1, 2

[16] K. Matzen and N. Snavely. Scene chronology. In *ECCV*. 2014. 2

[17] A. Mikulik, F. Radenović, O. Chum, and J. Matas. Efficient image detail mining. In *ACCV*, 2014. 2

[18] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007. 4

[19] G. Schindler and F. Dellaert. Probabilistic temporal inference on reconstructed 3D scenes. In *CVPR*, 2010. 2

[20] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 3

[21] J. L. Schönberger, F. Radenović, O. Chum, and J.-M. Frahm. From Single Image Query to Detailed 3D Reconstruction. In *CVPR*, 2015. 1, 2, 3, 4

[22] N. Snavely. *Scene reconstruction and visualization from internet photo collections*. PhD thesis, 2008. 2

[23] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. In *SIGGRAPH*, 2006. 2

[24] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *IJCV*, 2007. 2

[25] Y. Verdie, K. M. Yi, P. Fua, and V. Lepetit. TILDE: A Temporally Invariant Learned DEtector. In *CVPR*, 2015. 3, 4