Personal Photo Enhancement using Internet Photo Collections

Chenxi Zhang, Jizhou Gao, *Student Member, IEEE,* Oliver Wang, Pierre Georgel,
Ruigang Yang, *Senior Member, IEEE,* James Davis, Jan-Michael Frahm, Marc Pollefeys, *Fellow, IEEE*

Abstract—Given the growth of Internet photo collections we now have a visual index of all major cities and tourist sites in the world. However, it is still a difficult task to capture that perfect shot with your own camera when visiting these places, especially when your camera itself has limitations, such as a limited field of view. In this paper, we propose a framework to overcome the imperfections of personal photos of tourist sites using the rich information provided by large scale Internet photo collections. Our method deploys state-of-the-art techniques for constructing initial 3D models from photo collections. The same techniques are then used to register personal photos to these models, allowing us to augment personal 2D images with 3D information. This strong available scene prior allows us to address a number of traditionally challenging image enhancement techniques, and achieve high quality results using *simple* and *robust* algorithms. Specifically, we demonstrate automatic foreground segmentation, mono-to-stereo conversion, field of view expansion, photometric enhancement, and additionally automatic annotation with geo-location and tags. Our method clearly demonstrates some possible benefits of employing the rich information contained in on-line photo databases to efficiently enhance and augment one's own personal photos.

Index Terms—Image enhancement, internet photo collections, segmentation, 2D to 3D conversion, field of view expansion, photometric enhancement, geo-tagging and locating

1 Introduction

With the seminal work by Snavely et al [1] that used Internet photo collections (IPCs) for 3D reconstruction and visualization, many image editing operations have been developed to unlock the rich information contained in IPCs. Examples include colorization, image inpainting, and geo-tagging etc. In this paper, we present a comprehensive framework that uses IPCs to enhance one's *personal photo collections* (photographs generally containing unique individuals) taken at landmark locations. As demonstrated in Figure 1, by using the rich IPCs taken at Notre-Dame, a casual cellphone image suffering from over-exposure on the background and very limited field of view, can be dramatically enhanced or even turned into a stereoscopic image with minimal user interaction.

We expand upon the success of recent Internet-photo-based 3D scene reconstruction methods [2], [3] to automatically register one's photo to 3D models reconstructed from IPCs. Once registered, a regular photo is immediately augmented with 3D depth information and a rich set of registered images taken under varying illumination conditions. The extra information can lead to a variety of photo enhancement operations. In this paper we specifically demonstrate the following enhancement to personal photos taken at landmark locations:

- 2D to 3D conversion: A single image can be converted to a stereo pair. Compared to the typical stereo conversion process that simply segments the image into multiple layers, we demonstrate more vibrant stereo pairs with a full 3D model of the background.
- Expanding the field of view: The field of view of the original photo can be expanded dramatically. We use a simple method to select color for the expanded pixels to match the color appearance of the original image. Our simple method works remarkably well through leveraging the dense sampling of lighting conditions represented by the photos of the IPC.
- Photometric Enhancement: Color artifacts on the background (e.g., landmarks) can be easily fixed, including over/under exposure and glare. Using the rich illumination variations in images of the IPC, these operations can be done with minimum user intervention. Given that most cameras set exposure (and focus) on the face, being able to automatically fix the problems in the background is a handy feature.
- Annotation: In typical IPCs approximately 10% of the images are geo-located [3]. We use this information to automatically geo-locate novel photos. Our method extends the simple geo-location of models [4], [3] to provide increased accuracy by also using Google StreetView imagery, unregistered images, and temporal constraints. Our tag transfer uses a tag analysis to obtain the most

- *C.Zhang, J.Gao and R.Yang are with Center for Visualization and Virtual Environments, University of Kentucky, Lexington, KY, 40506. E-mail: chenxi.zhang@uky.edu, {jgao5, ryang}@cs.uky.edu.*
- *O.Wang is with Disney Research Zurich, Zurich, Switzerland, 8092. E-mail: owang@disneyresearch.com*
- *P.Georgel and J.Frahm are with Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599 E-mail: pierre.georgel@gmail.com, jmf@cs.unc.edu*
- *J.Davis is with Department of Computer Science, University of California, Santa Cruz, Santa Cruz, CA, 95064 E-mail: davis@cs.ucsc.edu*
- *M.Pollefeys is with ETH Zurich, Zurich, Switzerland, 8092 E-mail: marc.pollefeys@inf.ethz.ch*

Digital Object Indentifier 10.1109/TVCG.2013.77 1077-2626/13/$31.00 © 2013 IEEE

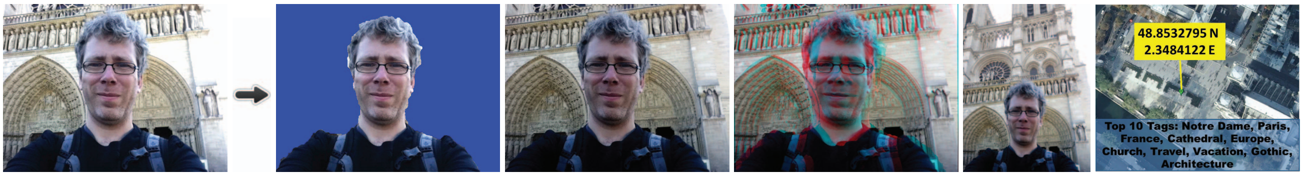

48.8532795 N
2.3484122 E
Top 10 Tags: Notre Dame, Paris, France, Cathedral, Europe, Church, Travel, Vacation, Gothic, Architecture

Fig. 1: Overview of our photo enhancements. From left to right: original image, foreground segmentation, photometric enhancement, stereoscopic image conversion, field of view expansion, and geo-tagging. The saturation (over-exposure) in right corner of the original image is removed by photometric enhancement, which is used as input for all the following enhancement applications. The field of view expansion result is intentionally resized for better figure layout. All photo enhancements are realized using Internet photo collections of the same landmark, Notre-Dame.

relevant tags for a particular view.

- Segmentation: we propose a novel foreground and background segmentation that automatically differentiates foreground based on image appearance statistic obtained from IPCs, surveying the scene over a variety of different illumination conditions. Compared to many existing segmentation methods, it requires no user interaction to obtain a high quality foreground segmentation. Even for complicated scenes with multiple foreground layers, tiny user interaction is only needed for separation of foreground layers.

While some of these effects have been demonstrated or mentioned previously in different contexts, our system is the first of using information from IPCs to solve these problems for personal photos. Our main observation is that with the abundance of variations in viewpoint and illumination in IPCs, quite often high-quality results can be obtained with relatively simple methods.

One main challenge we have to overcome is the limited completeness and accuracy found in state-of-the-art reconstruction methods due to occlusion and noise in the IPCs. An additional challenge when enhancing a personal photo is that we typically only have one view of the foreground in the personal photo. Therefore, its geometry can not be reconstructed. To overcome these limitations we propose novel methods to estimate or interpolate the missing depth values.

With the set of techniques we have adopted and developed in this paper, we envision a system where a user can upload his/her trip photos, taken with a regular camera, into a photo-editing system. The system will query previously reconstructed 3D models (from IPCs) and register each photo with the appropriate data set. After registration, geo-location as well as tags can be obtained, and a foreground and background segmentation can be automatically computed. Then, users can interactively enhance their trip photos in various ways. Our goal is to create powerful editing abilities with minimal interaction. Some methods operate automatically, but in certain cases some user input is required. We summarize these steps here, which will be clarified in detail in the following sections. We require user input to: (a) fit a ground plane, (b) assigning a depth level for the foreground layer in 2D-3D conversion, (c) isolate glare regions for photometric enhancement, (d) select from top candidates of homography warped images for FOV expansion to expand regions not covered by the 3D model (e.g, ground), and (e) isolate single foreground object from cluttered foregrounds. These operations only require a few mouse clicks each.

With more and more photos being stored online, we predict that such a system will with time become more and more robust to use, require less human interaction, and be able to accomplish an increasing number of applications over a large range of locations.

2 Related Work

There are three main fields of related work: Internet photo-based reconstruction, image enhancement, and image segmentation.

Internet photo-based reconstruction With the prevalence of consumer cameras and large scale on-line photograph storage sites, 3D modeling from IPCs has become a hot topic in recent years [1], [5], [2], [3]. Snavely et al. [1] presented pioneering work using photographs from IPCs to compute a 3D model reconstruction and recovered camera poses. Furukawa and Ponce [2] presented efficient clustering and filtering algorithms for parallel reconstruction that enforced inter-cluster consistency constraints over the entire reconstruction. Subsequently, Agarwal et al. [5] and Frahm et al. [3] advanced the state-of-the-art of city scale reconstruction from IPCs, with both improved geometric accuracy and computational performance. Our work adds onto these recent advancements to better perform incremental updates to these models, creating new sources of prior information for use in personal photo enhancement and augmentation.

Image enhancement There have been several techniques for image editing using large quantities of images downloaded from the web, such as colorizing gray-scale images [6], [7], enhancing CG images [8], and enhancing face images using good example prior photos of the same person [9]. Most related to our work is image completion [10], [11], [12], [13]. Although impressive results are presented, these completion methods focus only on image inpainting tasks, while our field of view expansion task is more general. Additional differences exist, for example, Hays and Efros [10] used semantically similar images for the completion task, but did not attempt to recreate the original scene. Whyte et al. [11] used photos of the same scene, but only applied homography corrections for the geometric registration of images. Our work on the other hand, combines 3D geometry information with homographies for more accurate image registration. Garg et al. [12] established a theoretical upper bound on the number of basis images to model real-world scenes and demonstrated some related applications including occluder removal and view expansion. However, their approach is limited when there are many large random and different foreground objects appearing in the images, requires manually segmenting a large set of images to learn the appearance bases, and is only able to output pure background landmark images with no foreground object appears. Instead, we introduce a novel use of content-aware scaling for challenging cases where there are many foreground occluders present. Our method is able to automatically segment foreground objects that exist in many personal photos, and create high quality field of view expansions in more general situations. [14] proposes a technique for intrinsic image extraction from photo collections and therefore can be used for lighting transfer between images. However, it cannot handle our saturation removal problems, since intrinsic image cannot be accurately extracted for saturated parts. Our proposed methods can solve the problem by replacing the saturated parts with content from properly exposed photographs in IPCs.

Our stereoscopic 3D creation application is inspired by work in the area of virtual view synthesis. View synthesis from multi-view data is a well established area. Most work involves computing depth from multi-view input, and using a depth-image based rendering (DIBR) model to create novel views [15], [16]. While our depth information is computed from multi-view data, it is all computed a-priori, and mapped to a single query image, which is used for view synthesis. In addition, so as to leverage sparse data, and avoid disocclusion problems, we use a robust image-domain warping method originally presented for artist-driven 2D to 3D conversion [17].

There are also works on estimating geographic information from images. [18] computes location distributions by low-level image matching to a geo-referenced database. [19] uses some travel priors to develop the chronological order of the images to find the location of images. Our geo-tagging method does not need such priors, and can obtain more precise geo-tags instead of geo-location distributions.

There has also been related work in the area of 3D model-based photo enhancement of landscapes and cityscapes [20]. Their work augments digital terrain and urban models with user interaction to register images to the 3D model, while we can achieve fully automatic 3D model selection and image registration. More significantly, we propose different enhancement applications from their work and our enhancement applications benefit not only from the reconstructed 3D models, but also from photographic appearance and other information that can be gained from large scale IPCs.

Image foreground segmentation Interactive image segmentation brings a user's prior knowledge of the location, size, color, and depth boundaries to segment a target object from an image, for example via a user-provided bounding box [21], [22] and strokes [23], [24]. However, even simple labeling tasks such as dragging a bounding box may still be daunting when dealing with lots of images. We leverage the opportunity that IPCs registered to the same model allow, enabling us to measure color consistency between the personal photo and IPC and filter out foreground and background color seeds, obtaining a high quality segmentation.

3 Overview

Figure 1 shows an overview of several challenging photo enhancements made easier by prior information gained from IPCs. Our approach assumes that IPCs of the relevant sites have been processed in advance so that geo-located 3D models of the relevant landmarks are available.

The proposed method starts by finding a set of 3D landmark models potentially associated to the personal image. We use an iconic scene graph based search [3] over the landmark models to identify a few potentially corresponding landmarks and their 3D information. Next, we identify the corresponding landmark through geometric verification by registering the personal photo with respect to the 3D model using SIFT feature matching and a RANSAC based robust pose estimation. A bundle adjustment refines the obtained registration of the personal photo. Then, a novel automatic foreground segmentation technique for separating occluding foreground objects from visible parts of the 3D model is used. After these pre-processing stages, we proceed to demonstrate four types of challenging enhancements to the personal photo: photometric enhancements(saturation and glare artifacts removal), stereoscopic image synthesis, field of view expansion, and geo-tagging, on seven different landmarks from man-made architecture to natural scene.

4 Modeling from IPCs

In order to verify the applicability of our proposed method, we apply two commonly used image-based reconstruction pipelines for testing. For IPCs that span a unique landmark, we reconstruct the camera locations using Bundler [1], which is an incremental structure-from-motion pipeline. We increase the density of the obtained point cloud by using PMVS [2]. For large collections that span several landmarks across a city, we use an iconic scene graph approach [3]. We refer to the resulting 3D point cloud as our *3D model* in future tasks. In order to offer a scalable solution, each image is represented by a binarized GIST descriptor [25] and the dataset is then clustered using K-medoids. Each of the obtained clusters is geometrically verified in a parallel fashion. Finally clusters are combined using hierarchical structure from motion.

We found that both of these solutions proved to be very successful on many IPCs. Non-rigid objects such as people in front landmarks are automatically ignored, and only rigid structures (such as the landmark itself) are reconstructed.

To fully exploit the reconstructed 3D model, a ground model is often necessary. Unfortunately, there are often too few reliably matched features on the ground to reconstruct an accurate 3D ground model from the photo collections. As a result, we design a simple interactive fitting tool that is used when no ground points are available in the 3D model. A RANSAC based automatic facade plane fitting is firstly applied on landmark, and then we specify the intersection line of this facade plane with the ground by selecting two points in the 2D image. The ground plane is derived by assuming that the facade plane and the ground plane are perpendicular in the real world. This assumption is valid in most man-made structures, and provides sufficiently accurate ground planes for our methods with little user interaction (only two clicks).

5 Personal Image Registration

Given a new image $\mathcal{Q}$ from one's personal photo collection (PPC), in order to perform image enhancement we first need to register $\mathcal{Q}$ to the reconstructed 3D landmark model $\mathcal{M}$. This can be considered as performing an incremental update of the reconstructed model.

In order to offer a scalable registration process for photos to an IPC, we propose a hierarchical matching approach, which first identifies a small set of potentially corresponding landmarks and then verifies registration to these landmarks. We use global image descriptors as shown in [3] to search for the k-nearest neighbors of image $\mathcal{Q}$ in the binarized GIST space [25]. This identifies a set of potential matching landmarks.

Then, a SIFT matching [26] between the candidate image $\mathcal{Q}$ and the collective SIFT descriptors of the 3D points of images registered within corresponding 3D model of each of the landmarks $\mathcal{M}$ is performed. To improve the robustness and the efficiency we used a mean-shift clustering [27] of the SIFT descriptors of the IPC model along the lines of Irschara et al. [28]. Next the images are registered into the IPC model using an efficient RANSAC [29] with a three-point registration [30]. This camera pose is further refined non-linearly to obtain the optimized camera pose of $\mathcal{Q}$.

While this method is efficient, it does not offer a satisfactory registration rate due to the fact that it only registers to views that contributed to the 3D models, and the surrounding areas of the landmark are often not represented in the 3D models. The key insight we use is that contrary to the un-ordered IPCs, personal photos are often a stream of images acquired during a tour of a city based on the referenced clock of a single camera. Hence the registration can be improved by jointly registering images from the PPC taken around the same time, as they are likely taken in close spatial proximity. Accordingly, for images that do not directly register with any model, we register them through robust fundamental matrix estimation, if they have overlap with registered images captured within a time range of up to an hour. Please note the choice of this range is not sensitive. If an overlap to a registered image in the PPC is found we use transitivity to propagate the registration. Examples of images registered using the time constraint are shown in Figure 2 for images taken over a time frame of 10 min.

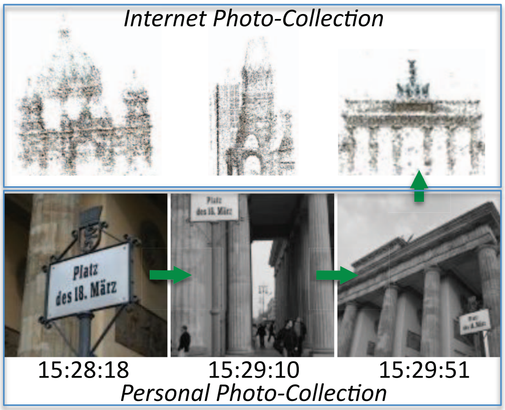


Fig. 2: For images that initially failed to register to a 3D model(first image in second row), we search for a match in the temporally neighboring images(second and third images in second row). After constraining their locations nearby, we are able to successfully register the input image to one of the reconstructed models(Among the three reconstructed models in first row, the input image is registered to the right one).

6 Image Foreground and Background Segmentation

One important prerequisite for many photo editing operations is the segmentation of the foreground. As opposed to interactive segmentation methods that rely on user interaction to learn the foreground and background appearance models, our method can acquire training data automatically based on the IPC. The fundamental assumption we make is that a pixel belonging to the background landmark is likely to be photometrically consistent across other views, whereas a foreground pixel usually is not.

Our method first projects the 3D model $\mathcal{M}$ onto the image $\mathcal{Q}$ denoted by m. The next step finds a set of images from the database that are captured at nearby locations under similar camera poses and image conditions of $\mathcal{Q}$, denoted by $\mathcal{S}$. Suppose a visible 2D point $p \in m$ is projected from the 3D point P and we denote its neighboring 3D point set as $N(P) = \{P' : \|P - P'\|_2 \leq 3 \cdot l\}$, where l is the average spacing between two closest 3D points. We then compute NCC(p, I_i), the normalized cross correlation of the color values of the projection of $N(P)$ on the image $\mathcal{Q}$ and the projection on an image $I_i \in \mathcal{S}$. We consider p is *consistent* between the image $\mathcal{Q}$ and the image I_i if NCC$(p, I_i) \geq 0.6$ or p is *inconsistent* if NCC$(p, I_i) \leq 0.2$. If NCC(p, I_i) is low because the projection of $N(P)$ lies on occlusion boundary, we still treat p as inconsistent between image $\mathcal{Q}$ and I_i. If p is consistent with the majority, i.e., over 80% of total number of images in $\mathcal{S}$, p is classified into the background seed set $\mathcal{B}$; similarly, if p is inconsistent with majority, p is classified into the foreground seed set $\mathcal{F}$.

We revise the initial setup of Grabcut [21] framework in two aspects: (1) we use automatically generated training data $\mathcal{F}$ and $\mathcal{B}$ to initially build the Gaussian Mixture Models for foreground and background instead of user-provided bounding box; (2) we add a constant penalty to the unary term of each pixel $p \in \mathcal{F}$ (or $p \in \mathcal{B}$) if p is labeled as background (or foreground) at the first run. We then perform the iterative energy minimization from Grabcut [21] to compute the segmentation. Figure 3 and Figure 5 compare fully automatic segmentation results from our method with Grabcut. Due to our precise color seeds used to train the appearance models for both foreground and background, our automatic approach achieves more accurate and meaningful segmentation than Grabcut.

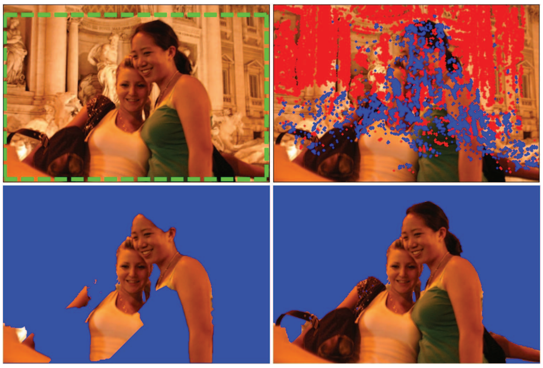


Fig. 3: Segmentation results: (top left) the original image superimposed with the bounding box prior for Grabcut, (top right) red dots for foreground color seeds and blue dots for background color seeds, (bottom left) the result of Grabcut, (bottom right) the result of our automatic approach.

7 Photo Enhancement

After the above pre-processing steps, we have a foreground-segmented image with a *detailed* 3D model. This rich representation allows us to perform a number of enhancements to the input image.

7.1 Stereoscopic Image Synthesis

Recently, stereoscopic 3D has seen a huge boom in popularity due to its ability of providing users with a more immersive viewing experience. Due to a combination of technological advances and the success of 3D movies in cinemas, home 3D displays have become increasingly commonplace. However, the production of personal stereoscopic content is still far from prevalent. Though Fujifilm stereo camera has started guidance for personal stereoscopic content creation, options for personal stereoscopic content creation are still limited, and general-case automatic 2D to 3D conversion is an unsolved and highly underconstrained problem. In this section, we propose a framework for generating a convincing stereoscopic pair from a single 2D personal photograph using prior information derived from large scale IPCs.

Depth Assignment In order to generate a stereo pair from a single image, we must first compute depth values for input image. We first apply the automatic foreground segmentation described in Section 5, and the depth values of background pixels are computed by projecting 3D model $\mathcal{M}$ to image plane. Note that sparse depth values are enough for our stereoscopic view synthesis described later.

Assigning depth values to foreground pixels however, usually requires user interaction. For images with a computed ground plane, we can assign the depth value for the foreground layer by back-projecting the ground contact point, e.g., one's feet, onto the 3D ground plane. However, for images without a visible ground plane, we allow the user to adjust its depth value interactively. In the most difficult case of multiple layers in the foreground, we developed a simple UI to allow further separation of the foreground. Our system supports strokes [24] and bounding boxes [21] to interactively separate different foreground objects. One such example is shown in Figure 5. It should be emphasized that this interactive step is only needed for images with complex foreground. Please see the supplemental material for an additional example of this interaction.

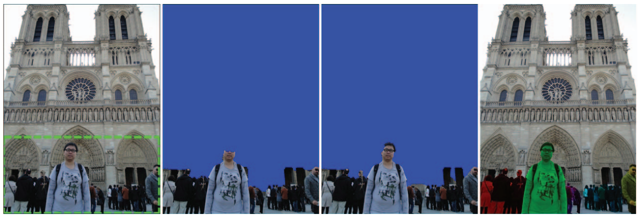


Fig. 5: Segmentation results for an image taken in a cluttered scene: from left to right, the original image superimposed with the bounding box prior for Grabcut, the result of Grabcut, the result of our automatic approach and interactively separated different (color coded) foreground objects.

After depth values for the scene are estimated, a virtual camera pose is computed such that the resulting stereo pair provides users with a natural and comfortable 3D viewing experience. It is indicated that for typical desktop displays with a viewing distance in the region of 700mm, the comfortable perceived depth range is 50mm in front and 60mm behind the display surface [31]. Similarly, we can

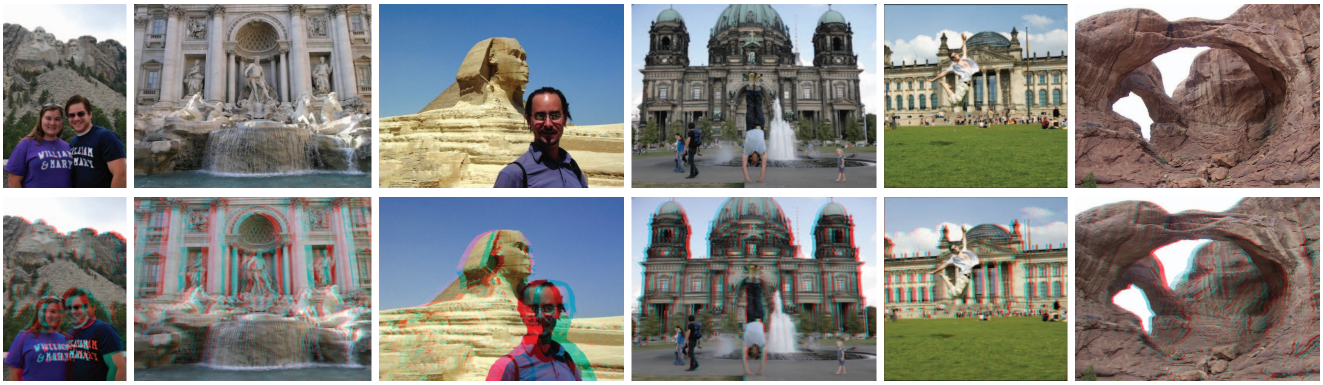


Fig. 4: Single-view personal photo to stereoscopic 3D conversion. Input images (top) are converted to stereo pairs (bottom) using depth information computed from IPCs. Note: As with any stereo 3D viewing situation, the perceived depth is related to the distance of the viewer to the screen. These images have been optimized for zoomed-in on screen display. Please view zoomed-in on screen for best results.

use the same method to retarget the photo to any other display. We compute a virtual camera pair with an optimal baseline such that the disparity range is mapped into this comfort zone of human visual perception using the mathematical derivation from the same work. We choose to synthesize both left and right views centered around the original image, rather than creating one virtual camera at twice the distance, as the reduced size of disocclusion regions lead to less artifacts and more convincing stereo results. Given the left and right camera poses and the depth map, we project pixels from known points into both views, computing a sparse set of disparities, which is used to synthesize the stereo pair.

Virtual view generation Stereoscopic view generation is a special case of the general virtual view synthesis problem, for which several classes of solutions exist. The most common of these is depth-image based rendering (DIBR), where a dense depth map is used to project each pixel into a novel view. However, these methods require per-pixel depth values, which can be difficult to compute, especially in untextured (sky) and unknown (disocclusion) regions.

Instead, we adopt a recent approach designed for 2D-3D conversion from scribble input [17] which we describe here for completeness. This method makes a piecewise continuous assumption that allows for discontinuities (determined automatically by our foreground segmentation) to appear at depth boundaries. To avoid disocclusions at these discontinuities, a two-step process is used. The first step computes a piecewise continuous image warp driven by our sparse disparity constraints. The second step stretches the background, using a content and disparity-aware retargeting method to fill in any disocclusions that have arisen. This method allows us to automatically generate high quality stereo pairs using our IPC computed depth prior.

Figure 4 shows some results of our stereo view generation. We also provide a comparison showing the naive approach of using a simple planar background after segmenting the foreground, as shown in Figure 6. We can see that our model $\mathcal{M}$ provides a more realistic and convincing depth impression by giving shape to the background regions. In addition, in cases where no foreground exists, we can still achieve a compelling stereoscopic image.

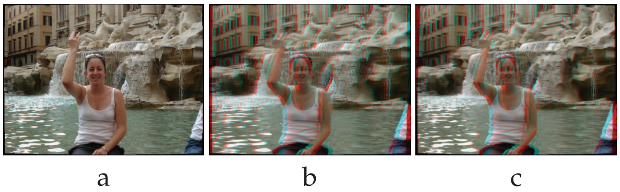


Fig. 6: Comparison showing the effect of incorporating our 3D model $\mathcal{M}$ into the stereoscopic conversion. a) Original image, b) result computed using a planar approximation to the background, c) result computed using $\mathcal{M}$. In the latter case, the rocks on the right can be seen at their correct depth level. Please view zoomed-in on screen for best results.

7.2 Field Of View Expansion

A common problem in photos is the limited field of view (FOV). This is particularly pronounced in self-portraits, such as the one shown in Figure 1. Here we discuss our approach to expand the original image's FOV using the background model from an IPC. Compared to the stereoscopic synthesis, the expanded image can contain a significant amount of missing data. Therefore a different synthesis method is presented.

Geometric registration The first step for FOV expansion is to warp similar images from the IPC to the query image's camera pose. The expanded region is then filled in mainly by the content from warped images, as well as texture synthesized using repetitive content.

Specifically, for pixels that exist in the 3D model $\mathcal{M}$, we use forward warping, projecting these points to the pose of the query image. While this accounts for a majority of registered pixels, projecting 3D point clouds can lead to holes in the image. In order to fix these small holes, we use a bilateral interpolation on the depth map prior to projection, which interpolates missing depth values weighted by color similarity and

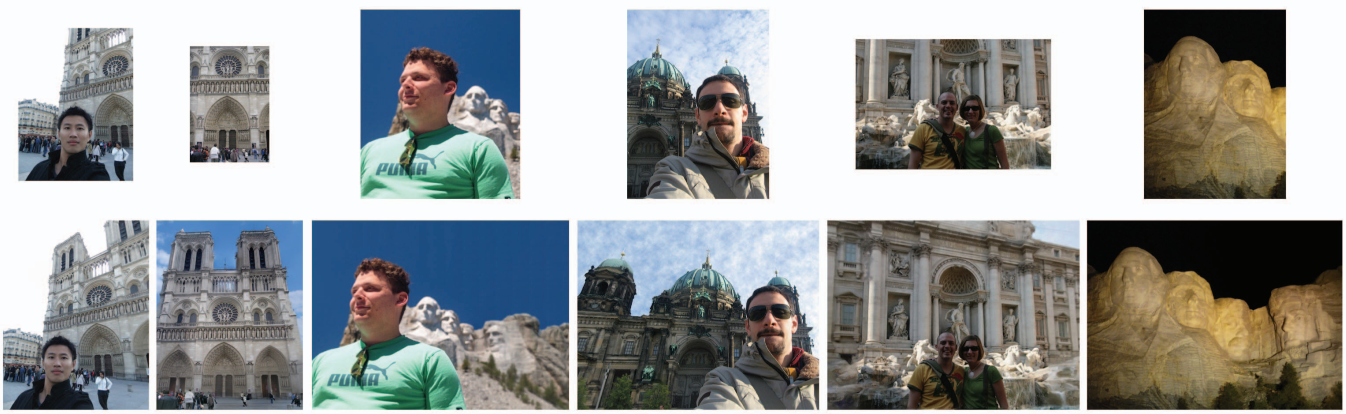


Fig. 7: Results from FOV expansion, (top) original images (bottom) FOV expanded images. For the third set, the expanded background is blurred with a low-pass filter to match the input.

spatial distance measurements. For parts of the image that do not appear in the 3D model, mostly sky or ground in our scenes, we mainly use texture synthesis or homography warping from nearby images to fill in information. Because of the diversity of lighting conditions and foreground objects in IPCs, it is necessary to select images with high color similarity to use in FOV expansion. In our work, we use a SSD metric to measure color similarity of the reference image and the warped source images. The top 20–50 images with highest color similarity are automatically selected for blending.

We choose a simple scheme for blending the selected images, using the median color of the top ranking matches, projected into the camera pose of the reference image. Figure 8 shows the quality of the the median color images after forward warping, and also the regions which we are not able to model using the geometry. These areas are then filled either by texture synthesis or by warping nearby images with a homography.

Using nearby images Due to the power of large scale IPCs (or the similarity of people's vacation photos), we found that often times we have numerous images from nearby camera poses. When selecting from these images, we favor images that have not only similar camera location and orientation, but also a wider field of view than image $\mathcal{Q}$, as these images will exhibit less distortion after warping, and will have a higher chance of containing the content needed to fill out our view. From these images, the top candidates are automatically chosen for blending the median image. Thanks to the large number of images in our database, we are often able to automatically find many images taken with very similar poses, which makes forward warping fairly accurate. However, for cases where there are no nearby images with the similar perspective, or when the query image has much higher resolution than source images, the forward warping naturally leads to blur in the filled image region (see Figure 18).

Combining sources Once we have geometrically and photometrically registered the images, the next step is to fill the expanded areas. To achieve seamless blending of different registered images, we combine gradients from all the sources, forming a new gradient image with expanded FOV. The output gradients are automatically combined, and the priorities are: original image (1st), geometrically registered median image (2nd), homography warped nearby images (3rd) if we have to use(such as ground). Specifically, to compose gradients of expanded FOV, we keep gradients from original image, and then add in gradients from median image to expanded area. For regions that median image does not cover, we fill in gradients from homography warped images. After we compute our combined gradients map, we solve Poisson's equation [32] to seamlessly reconstruct the output image. Texture synthesis will be applied if sky region is missing, as will be described in next paragraph. When the background is out of focus, we allow users to specify a Gaussian blur for the expanded region, to match the reference image(column 3 in Figure 7). Figure 7 shows some results from our FOV expansion application.

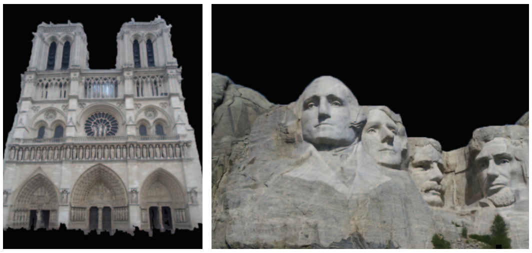


Fig. 8: Median color images from selected geometrically registered images with high photometric consistency.

Content-aware scaling and texture synthesis There are some extremely challenging cases where important occluders, such as people, are cropped by the image border. In this case there is no way to reconstruct the remaining portions of these occluders with our model $\mathcal{M}$ *or* with nearby images. We present a solution to this problem where we first apply automatic segmentation to extract the important foreground objects, and then use a content-aware image resizing technique [33] to stretch the remaining unimportant background regions(such as ground), while preserving the aspect ratio of foreground objects. This is used to *expand* the background, without creating artifacts due to foreground objects being cut off at the borders. As shown in the second image of Figure 7, by our novel usage of seam carving allows a photo-realistic expansion of original image to be obtained for a difficult case.

Finally, as mentioned above, for sky regions (segmented using our geometric model) that remain unfilled after all prior compositing steps, we implement a texture synthesis approach [34] to fill the remaining holes. When there is no sky region at all in the original image, we use an infinite homography to warp the sky from the other nearby images, and then complete the region with texture synthesis. By combining a number of simple approaches, our system is able to achieve high quality results, filling in convincing information from a large set of registered images with minimal user interaction.

7.3 Photometric Enhancement

IPCs also provides an excellent sample set to fix up problematic areas in personal photos, in particular these areas on the background (e.g., the landmark). It should be noted that most cameras determine metering based on the central content, or even on detected faces. Therefore it is more likely that the background part requires photometric correction.

Flawed area identification In this application we mainly deal with two types of enhancements, removing saturated regions and glare. We automatically detect saturated regions (over-exposed and under-exposed) by simply thresholding r,g,b values of pixels near 0 and 255. Automatic detection of glare is more difficult. Therefore we allow a user to identify the location of these artifacts, as shown in Figure 10.

Image composition We adopt the same scheme as in Sec 7.2 for computing a median color image over the registered images with high photometric consistency, using SSD to measure the color similarity of reference image and the warped source images (excluding areas that needs to be enhanced). Under this scheme, we find a median color image that contains important image detail within these saturated and glare regions, while maintaining color similarity to the reference image. To achieve seamless blending, we again perform Poisson blending [32], replacing gradients of the saturated regions in the personal photograph with those from median image.

Tone mapping Introducing detail into the saturated region can result in an HDR image whose dynamic range is beyond the 8-bits of the input image/display device. Therefore, as a post-processing step, we apply a standard tone mapping technique to obtain the final, viewable image. Figure 9 and Figure 10 shows some results of our photometric enhancement application and comparison with results from Photoshop.

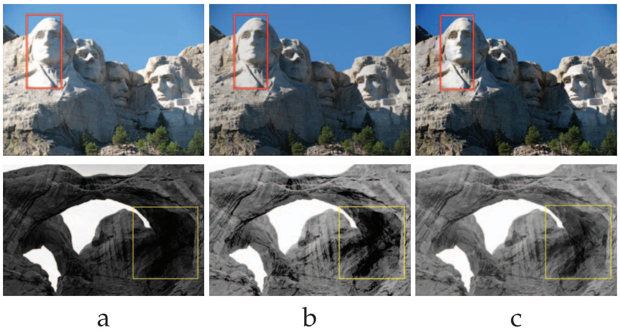


Fig. 9: Comparison showing the effect of our saturation removal using internet photo collections. a) Original image, b) Result from Photoshop color adjustment, c) Result from our method. Since saturated regions have no details, simple color adjustment cannot fix the problem. Please view zoomed-in on screen for best results.

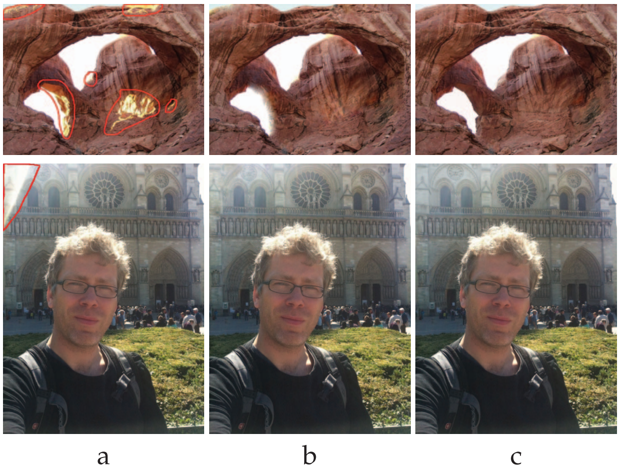


Fig. 10: Effect of our glare removal using internet photo collections. a) Original image with user selected glare region, b) Results using Photoshop Healing and Clone tools, c) Results from our method. Please view zoomed-in on screen for best results.

7.4 Transfer of Information

IPCs are not limited to a set of images, they typically also incorporate meta-data such as geo-location and text tags. When a PPC is aligned to an IPC, these meta-data can be automatically transferred. This offers additional ways to browse your PPC either by displaying your images on a map or for example by grouping the images by label. In the following section, we describe how our system transfers geo-locations and text tags from IPCs to PPCs.

Geo-tagging While geo-location of the PPC photos is often desired for visualization effects (map of location of a collection, browsing, etc.) it is a tedious manual task to localize photos without GPS information. We find that in an IPC typically more than 99% of the photos do not contain GPS information. Our system removes the burden of manual localization from the user by automatically geo-localizing the photos of the PPC. We use the embedded tags (automatic and user clicked) of the images in the IPC. In order to support more accurate geo-location we also use Google StreetView panoramas, which have high accuracy geo-location and orientation[1].

First we translate the GPS latitude longitude information into an approximate local metric coordinate system the Universal Transverse Mercator (UTM) grid. This conversion allows us to consider Euclidean distances between geo-tags.

To obtain geo-localization for the PPC, our system uses a two-stage process. In a first preprocessing stage executed once for each model in the IPC the system obtains an accurate geo-localization of the corresponding IPC model. This accurately localized model is then used to perform a geo-localization of the PPC images.

IPC model localization: Our algorithm first obtains an approximate geo-localization through a kernel voting. Each image that has a geo-tag votes for its location through a Gaussian kernel centered at its geo-tag location and a 3σ cut off distance of 25 meters for clicked geo-tags (approximate clicking accuracy) or a 3σ cut off distance of 10 meters for GPS based geo-tags. To suppress outliers we then select all geo-locations within the biggest mode as the set of valid geo-tags. The approximate location of each model is then obtained by the averaging of the inlier locations. Alternatively, when there are no geo-tags or no reliable geo-tags (no consistent votes) we use the text-tags of the images of the IPC for a location query on Google Maps to obtain an approximate geo-location.

The approximate geo-location of the IPC model is then used to obtain all nearby Google StreetView imagery (panoramas) available for the refined geo-localization, whose location is likewise transferred into the UTM coordinate system. Given that these images contain mostly road surfaces and cars below the horizon line we discard all information below the horizon for the further processing, (the horizon can be directly obtained from the image orientation). All panoramas are then registered into the model using our registration process from Section 5 but instead of using the standard three-point algorithm we use a three-point algorithm based on viewing rays given that the panorama directly provides viewing rays. Then using a RANSAC approach we transform the IPC model coordinate system into the UTM coordinate system. We use the known positions of the panoramas in the IPC model and the UTM coordinate system of the same panorama as correspondences to estimate the transformation from the IPC model coordinate system to the UTM coordinate system. This step delivers an accurate transformation from the IPC model coordinate system to the geo-coordinate system. We apply this transformation to geo-localize the IPC model, which is then be used in the next step to geo-localize the images of the PPC.

PPC image geo-loclization: This is using the registration process discussed in Section 5 with respect to the geo-located model. This directly obtains geo-location in the UTM coordinate system for the photos of the PPC.

Typically there is a large fraction of the IPC images with geo-tags that are not registered with our model. These images provide valuable geographic information about less popular scenes in an IPC. Therefore for the images that failed to register to a geo-localized model, we search for matching images in the set of geo-tagged images using the same method as in Section 5 including trying to match images of the PPC taken at approximately the same time. The obtained geo-tags are then filtered using the same kernel voting as the IPC model geo-location. Matching to the complete set of geo-tagged images from the IPC drastically increases the number of geo-localizable images in the PPC.

We tested this geo-tagging approach on several PPCs that were taken in Berlin by three different users. We used an IPC consisting of 2.8 millions images retrieved from Flickr (including 353,584 geo-tagged images) and 467 geo-localized 3D models. In this case we were able to geo-tag more than 55% of the input images with an estimated accuracy of approximately 50 meters, which is related to the quality of input geo-tagged information. Visual geo-tagging results are shown in Figure 11.

Image labeling Labeling images offers additional information to ones PPC. It not only allows the user to search images in his or her PPC based on keywords, but also to retrieve additional information about a photo. For example, the tags found for an image of a monument are usually precise enough for Google to retrieve the corresponding Wikipedia article.

In order to offer precise tag candidates to the user, we propose a hierarchical automatic annotation algorithm. First we select the most popular tags from the complete IPC from which the model was computed. Then for each image we select the most popular tags from the particular landmark it was registered to, excluding the previously attached tags from the IPC. If an image is registered to several landmarks, we select the tags that have the highest combined mean occurrence across the landmarks (number of times a given tag is represented in the dataset divided by the total number of tags). Finally we add tags coming directly from the images the candidate image registered to, adding a few more localized tags.

In addition to these effects, we could also easily perform other enhancements based on the scene depth, such as refocus, depth-of-field control, etc. These effects have been demonstrated with depth obtained by other means, therefore we will not show examples in this paper.

[1] The panoramas used are automatically downloaded through the Google StreetView API
http://code.google.com/apis/maps/index.html

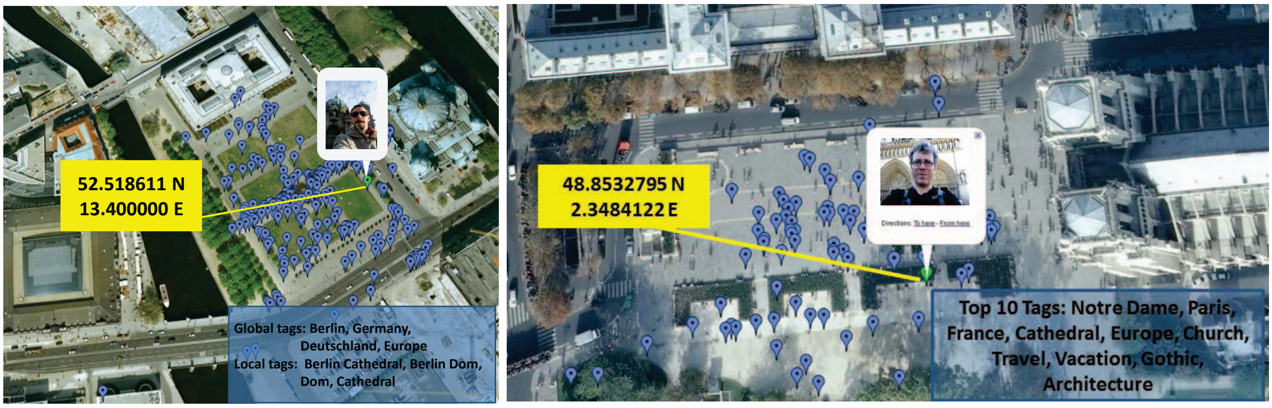

52.518611 N
13.400000 E
Global tags: Berlin, Germany, Deutschland, Europe
Local tags: Berlin Cathedral, Berlin Dom, Dom, Cathedral
48.8532795 N
2.3484122 E
Directions: To here - From here
Top 10 Tags: Notre Dame, Paris, France, Cathedral, Europe, Church, Travel, Vacation, Gothic, Architecture

Fig. 11: Geo- and textual tagging of personal photos using IPCs of Berlin and Notre Dame. Locations depicted with green pins correspond to the query images that are matched a 3D model, geo-tagged and geo-localized. Blue pins represent images from IPC.

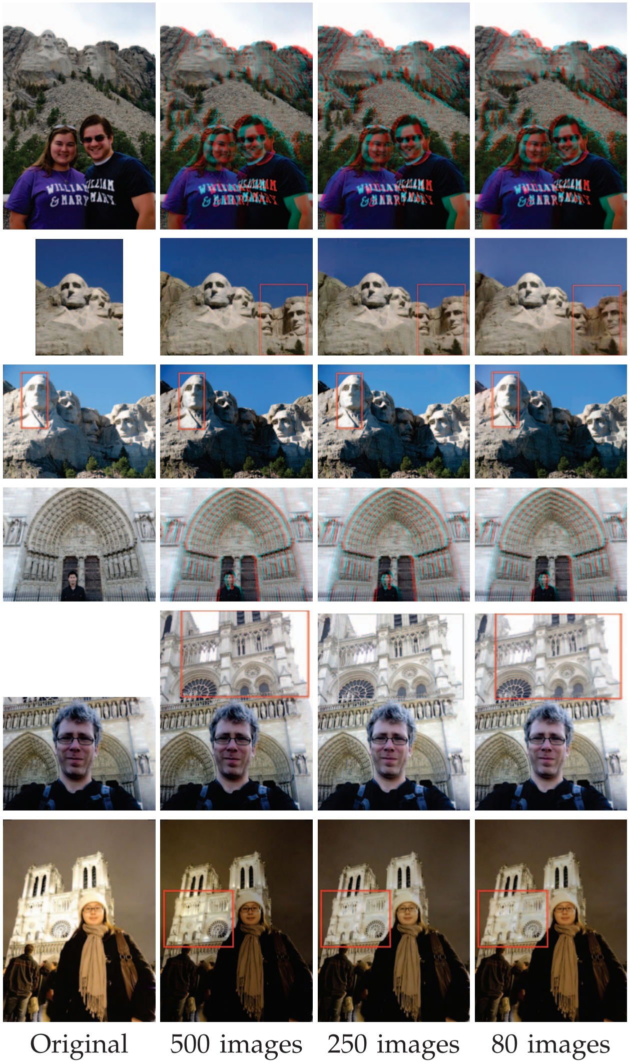


Fig. 12: Comparison showing how the results depend on the number of images in database. Row 1-3: Mount Rushmore Dataset. Row 4-6: Notre Dame Dataset. From top to bottom: 2D to 3D conversion, FOV expansion, photometric enhancement. Please view zoomed-in on screen for best results.

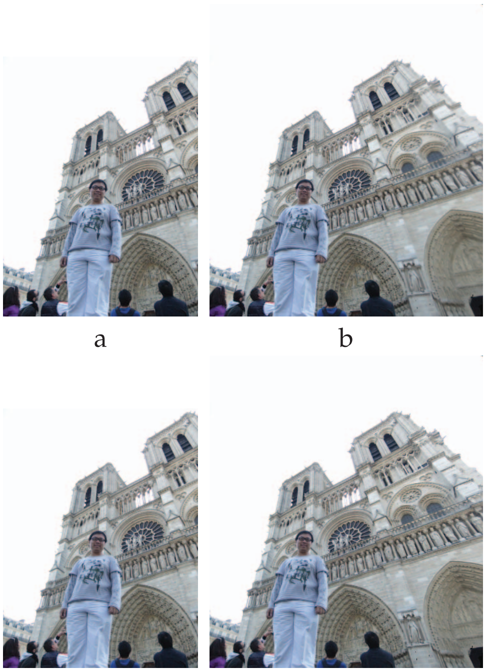
8 Evaluation

To further evaluate the performance of our system, two groups of evaluations are conducted. The first one evaluates the operating range of our methods, such as how the results depend on the number of images in database, or the resolution/quality of the query image. The second evaluates how convincing users find the enhancement results, as well as our system's processing speed and interaction requirements, especially by comparing with state-of-the-art image editing software.

8.1 Operating Range Evaluation

The more images the database contains, the richer information we can obtain. In this evaluation, we evaluate how our enhancement results depend on the number of images in database. We choose the Mount Rushmore dataset to demonstrate it. We carried out the same operations, varying only the number of images in database - 500 images, 250 images, 80 images. As shown in Figure 12, the first row demonstrate that with more images in database, the 3D effects are more realistic. In the result using a 500 image database, we can see more 3D geometry variance in the far rock, which disappears when only using 250 images. When only 80 images are used, the depth effect becomes unnatural in the background. Those differences are due to the density of the reconstructed 3D point clouds. The second and third rows show that more images induce more realistic FOV expansion and photometric enhancement, due to a larger number of photometrically-consistent images that can be drawn from. This can be seen especially in the area between the two rightmost faces in FOV expansion case, and the leftmost face in photometric enhancement case. Similar comparison can be seen in another dataset(Row 4-6 in Figure 12).

Another factor that affects our system's performance is the resolution/quality of the query image, especially on FOV expansion application. If the query image is of high resolution compared to other images in database, the geometric warping from database images to reference image will be more blurry, therefore producing a more blurry median image and final FOV expanded image. Conversely, if the query image is lower resolution than other images in database, the warped images can appear sharper than the query image. Generally, the more consistent resolution/quality a query image has with images in database, the better performance that our system can achieve. As shown in Figure 13, an image is expanded at two different resolutions, using the same IPC. Although the overall resolution of (b) is much higher than (d), the expanded region of (d) is more consistent to the quality of original image(c), therefore looks more natural, while in (b) the expanded region is more blurry relative to the quality of original image(a), which looks a little bit unnatural.


a b

c (downsampled) d (downsampled)

Fig. 13: Performance related to query image resolution. a) Original image, b) Result from original image, c) Downsampled original image(X5), d) Result from downsampled(X5) original image. The expanded region of (d) is more consistent to the quality of (c), therefore looks more natural, while the expanded region of (b) is more blurry relative to the quality of (a), which looks a little bit unnatural. Please view zoomed-in on screen for best results.

8.2 Performance Evaluation

We conducted a user study to evaluate how realistic our photo enhancement results are compared to the processing results from state-of-the-art image editing tools, e.g., Photoshop. Four Photoshop experts were asked to process 20 query images from 6 different scenes, among which 4 images are used for segmentation, 6 images for 2D to 3D conversion, 4 images for FOV expansion, 6 images for photometric enhancement.

Based on the design goal of our system, we hypothesize that using our system will be able to complete the tasks significantly quicker than using Photoshop, the state-of-the-art image editing tool. The average time required to complete the four tasks with our system and Photoshop are illustrated in Figure 14. We can see that for all the four tasks, our system takes much less time than using state-of-the-art image editing tools. ANOVA tests confirm that the time differences are statistical significant for all the four tasks($F = 10.87, p\text{-}value = 0.008 < 0.01$ for "Foreground Segmentation"; $F = 30.42, p\text{-}value = 0.0002 < 0.01$ for "2D to 3D Conversion"; $F = 17.21, p\text{-}value = 0.006 < 0.01$ for "FOV Expansion"; $F = 47.04, p\text{-}value = 0.002 < 0.01$ for "Photometric Enhancement"), which further validates our hypothesis.

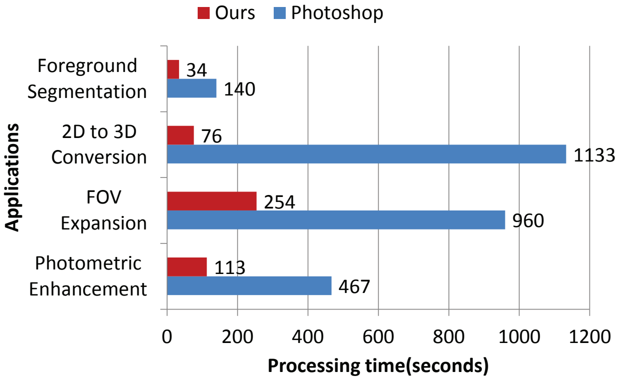
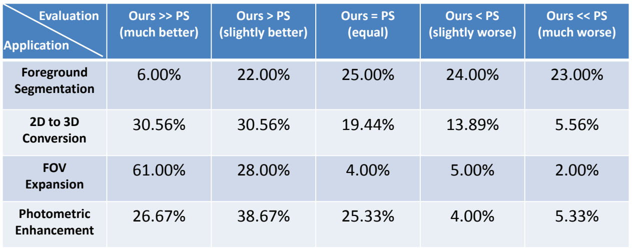


Fig. 14: Average time used for processing under our system and Photoshop.

25 users are asked to compare the Photoshop results with our enhancements on the same images. To avoid bias, images from two methods are randomly ordered and users do not know which image comes from which method. Users were required to select one of five preference choices: 1) Image 1 is much better than Image 2; 2) Image 1 is slightly better than Image 2; 3) Image 1 is equal to Image 2; 4) Image 1 is slightly worse than Image 2; 5) Image 1 is much worse than Image 2.



Evaluation / Application	Ours >> PS (much better)	Ours > PS (slightly better)	Ours = PS (equal)	Ours < PS (slightly worse)	Ours << PS (much worse)
Foreground Segmentation	6.00%	22.00%	25.00%	24.00%	23.00%
2D to 3D Conversion	30.56%	30.56%	19.44%	13.89%	5.56%
FOV Expansion	61.00%	28.00%	4.00%	5.00%	2.00%
Photometric Enhancement	26.67%	38.67%	25.33%	4.00%	5.33%

Fig. 15: User study result: Percentage of preference choices between our results and Photoshop(PS) results. The four rows represent four different applications. The expert-driven PS segmentation can be deemed as ground truth. Users' visual evaluations show that our segmentation results are comparable to ground truth. For the other three applications, our results receive much higher evaluation scores.

As shown in Figure 15, the user study clearly shows that our results are either favored over those from Photoshop, or comparable (for the difficult task of automatic foreground segmentation). It is worthwhile to point out that taking Photoshop segmentation as ground truth, our automatic foreground segmentation achieves an average error rate of 3%, with minimum error rate of 0.7% and maximum error rate of 8.7%(multiple foreground layers case). This indicates that our approach is capable of automatically producing comparable foreground segmentations to hand-tuned foreground segmentation maps in Photoshop.

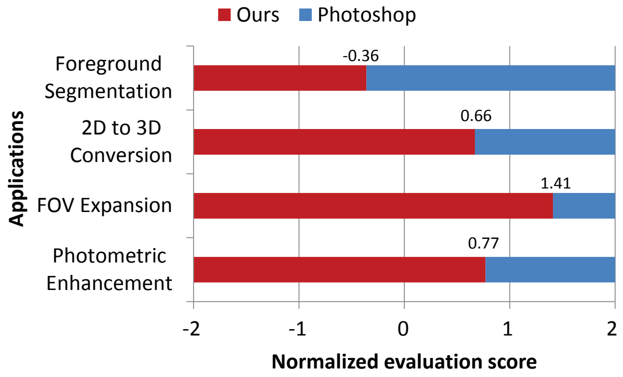


Fig. 16: Normalized evaluations of Our approach vs Photoshop.

The error rate is computed as the percentage of mislabeled pixels.

Figure 16 shows a more intuitive diagram of user preference. We normalize the five statements in Figure 15 to numerical scores ranging from +2 to -2. The mean value of these scores lies on the boundary between red bar(ours) and blue bar(Photoshop). The larger the normalized evaluation score, the more preferable our approach was to Photoshop. Figure 17 shows some visual comparisons between our approach and the Photoshop images that we used in the study. This user study validates our conclusions again that leveraging IPCs and simple user interaction allows us to create more convincing image enhancements than what is possible with state-of-the-art image editing tools used by experts.

In terms of geo-tagging and locating evaluation, unfortunately we are not aware of a system that tries to geo-localize PPC to IPCs. Most systems only register images to a geo-registered model. In that our approach is similar when we can register the input image to a model, but when this fail we offer a backup to transfer geo-tag which, as far as we know, was not proposed before. Therefore we believe that these results are relevant. As for the geo-registering of model we offer a robust system based on commodity panoramas (e.g. Google street view images). Current system uses accurate GPS or a combination geo-tags and maps. We currently do not compare to this system, but our attempts to geo-register our model using only the geo-tags have failed because of the amount of noise in the input.

9 Discussions

Like many IPC-based approaches, the success of our approach depends on the availability of large sets of photos taken at the same site. Therefore its application is currently mainly limited to images of touristic sites. However, there are on-going efforts for large-scale image acquisition through competitive games [35]. So we are optimistic that the applicability of our approach will be greatly expanded in the future.

Fully automatic image segmentation is a challenging task. As shown in Figure 5, our automatic approach sometime is not sufficient to further separate foreground layers for 2D-3D conversion. Therefore we still require user interaction to separate the foreground layers. This is the most time-consuming part of our entire system (for which we have prepared a video in the supplementary materials).

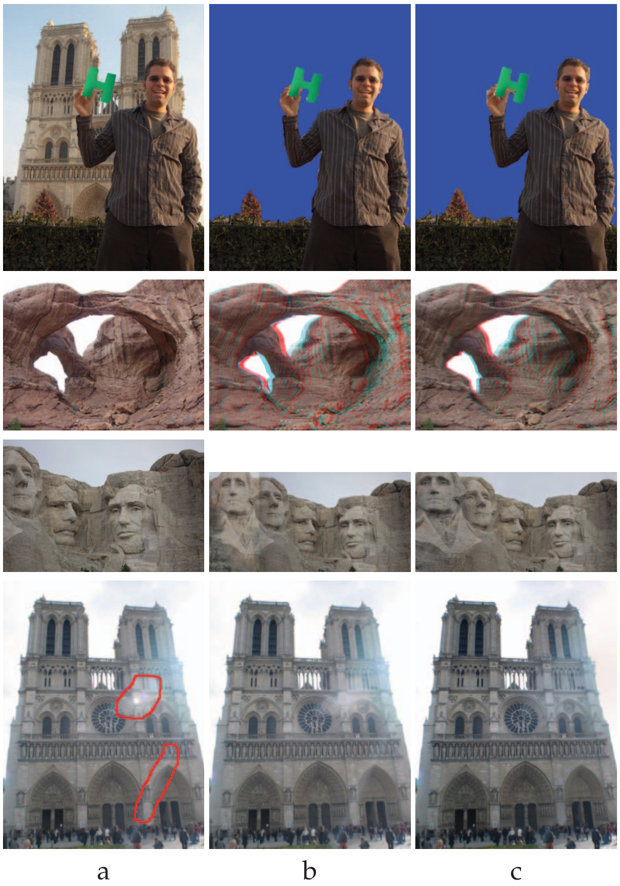


Fig. 17: Comparison showing the effect of our enhancements using internet photo collections. a) Original image, b) Result from Photoshop, c) Result from our method. From top to bottom: foreground segmentation, 2D to 3D conversion, FOV expansion and photometric enhancement. Please view zoomed-in on screen for best results. To see more photometric enhancement comparisons with Photoshop, please refer to Figure 9 and Figure 10. More comparison results are provided in supplemental materials.

FOV expansion is also a difficult case, especially for areas that have no 3D information (Figure 18(a)) or inaccurate 3D information (Figure 18(b)). One example of such areas, are locations that are non-static, such as a fountain, which cannot be accurately reconstructed by our 3D model. This causes the corresponding location to become blurry when reconstructed from median color values of nearby IPC images (Figure 18(b)). One solution to this problem could be by including manual interaction to select patches where the result is sampled from a single image.

Generally, our proposed enhancement methods work on both daytime and nighttime images, however, personal image registration is more challenging for nighttime image. As long as there exist photometrically-consistent photographs to input image in database, the gradient field of created median image is good to use for our proposed enhancements.

If the input image is taken under too extreme condition(e.g sunset or glossy scenes), there is high possibility that very few photos under similar condition can be retrieved from database, the proposed method might not work quite well. Another limitation is that our median image based method might result in inconsistent lighting, since we didn't take into account the lighting direction when creating median image, especially when there is strong directional lighting in the input image.

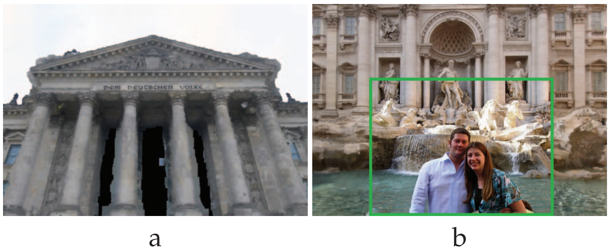


Fig. 18: Limitations of our approach: (a) Missing data behind the columns caused bad interpolated depth values, and consequently a low-fidelity synthesized view in the column areas. (b) Another example of field-of-view expansion. The green box is the original photo. Note that the flowing water from the fountain on both the left and right sides, and the fountain rocks are blurry.

10 Conclusion

In this paper, we approach personal photo enhancement from a novel direction - using IPCs. Our work leverages the 3D background models reconstructed from IPCs of the same landmark. With the rich information from large scale IPCs, we believe that by augmenting one's personal photo with depth information, as well as the surrounding appearance information, a number of interesting photo enhancements can be achieved. Applications that we have explored vary from automatic image segmentation, stereoscopic view synthesis, to field of view expansion, photometric enhancement, geo-tagging etc, all of which show promising results and validate the potential of our approach.

Although the 3D model is reconstructed from internet photos using the state-of-the-art techniques, the sparsity and inaccuracy of the 3D model still contribute to some failure cases in photo enhancements. Newer approaches like Frahm et al. [3] provide dense geometry which could help overcome the sparsity limitations. The 3D depth augmentation of 2D images also enables additional enhancements such as super-resolution or 3D re-targeting by using the full variety in resolution and illumination conditions captured by the IPC.

Finally, in this paper we only focus on single image enhancement. However one future direction could be to extend our work to multiple images or even video sequences for enhancement.

References

[1] N. Snavely, S. M. Seitz, and R. Szeliski, “Photo tourism: exploring photo collections in 3d,” *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, 2006.

[2] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, “Towards internet-scale multi-view stereo,” in *CVPR*, 2010, pp. 1434–1441.

[3] J.-M. Frahm, P. F. Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, and S. Lazebnik, “Building rome on a cloudless day,” in *ECCV (4)*, 2010, pp. 368–381.

[4] R. Kaminsky, N. Snavely, S. M. Seitz, and R. Szeliski, “Alignment of 3d point clouds to overhead images,” in *Workshop on Internet Vision in conjunction with CVPR*, 2009.

[5] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, “Building rome in a day,” in *ICCV*, 2009, pp. 72–79.

[6] X. Liu, L. Wan, Y. Qu, T.-T. Wong, S. Lin, C.-S. Leung, and P.-A. Heng, “Intrinsic colorization,” *ACM Transactions on Graphics (SIGGRAPH Asia 2008 issue)*, vol. 27, no. 5, pp. 152:1–152:9, December 2008.

[7] Y. S. Chia, S. Zhuo, R. K. Gupta, Y.-W. Tai, S.-Y. Cho, P. Tan, and S. Lin, “Semantic colorization with internet images,” *ACM Transaction on Graphics(TOG)*, 2011.

[8] M. K. Johnson, K. Dale, S. Avidan, H. Pfister, W. T. Freeman, and W. Matusik, “Cg2real: Improving the realism of computer generated images using a large collection of photographs,” *IEEE Transactions on Visualization and Computer Graphics*, 2011.

[9] N. Joshi, W. Matusik, E. Adelson, and D. Kriegman, “Personal photo enhancement using example images,” *ACM Trans. Graph*, vol. 29, no. 2, pp. 1–15, 2010.

[10] J. Hays and A. A. Efros, “Scene completion using millions of photographs,” *Commun. ACM*, vol. 51, no. 10, pp. 87–94, 2008.

[11] O. Whyte, J. Sivic, and A. Zisserman, “Get out of my picture! internet-based inpainting,” in *BMVC*, 2009.

[12] R. Garg, H. Du, S. M. Seitz, and N. Snavely, “The dimensionality of scene appearance,” in *ICCV*, 2009, pp. 1917–1924.

[13] K. Dale, M. K. Johnson, K. Sunkavalli, W. Matusik, and H. Pfister, “Image restoration using online photo collections,” in *International Conference on Computer Vision*, 2009, pp. 2217–2224.

[14] P.-Y. Laffont, A. Bousseau, S. Paris, F. Durand, and G. Drettakis, “Coherent intrinsic images from photo collections.”

[15] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. A. J. Winder, and R. Szeliski, “High-quality video view interpolation using a layered representation,” *ACM Trans. Graph.*, vol. 23, no. 3, pp. 600–608, 2004.

[16] L. Zhang, W. J. Tam, and D. Wang, “Stereoscopic image generation based on depth images,” in *ICIP*, 2004, pp. 2993–2996.

[17] O. Wang, M. Lang, M. Frei, A. Hornung, A. Smolic, and M. H. Gross, “Stereobrush: Interactive 2d to 3d conversion using discontinuous warps,” in *SBM*, 2011, pp. 47–54.

[18] J. Hays and A. A. Efros, “Im2gps: estimating geographic information from a single image,” in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[19] E. Kalogerakis, O. Vesselova, J. Hays, A. A. Efros, and A. Hertzmann, “Image sequence geolocation with human travel priors,” in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 253–260.

[20] J. Kopf, B. Neubert, B. Chen, M. F. Cohen, D. Cohen-Or, O. Deussen, M. Uyttendaele, and D. Lischinski, “Deep photo: model-based photograph enhancement and viewing,” *ACM Trans. Graph.*, vol. 27, no. 5, p. 116, 2008.

[21] C. Rother, V. Kolmogorov, and A. Blake, “”grabcut”: interactive foreground extraction using iterated graph cuts,” *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

[22] V. S. Lempitsky, P. Kohli, C. Rother, and T. Sharp, “Image segmentation with a bounding box prior,” in *ICCV*, 2009, pp. 277–284.

[23] Y. Boykov and M.-P. Jolly, “Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images,” in *ICCV*, 2001, pp. 105–112.

[24] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum, “Lazy snapping,” *ACM Trans. Graph.*, vol. 23, no. 3, pp. 303–308, 2004.

[25] M. Raginsky and S. Lazebnik, “Locality Sensitive Binary Codes from Shift-Invariant Kernels,” in *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[26] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[27] D. Comaniciu, P. Meer, and S. Member, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 603–619, 2002.

[28] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, Mar 2009.

[29] R. Raguram, J.-M. Frahm, and M. Pollefeys, "A comparative analysis of ransac techniques leading to adaptive real-time random sample consensus," in *ECCV (2)*, 2008, pp. 500–513.

[30] R. M. Haralick, C. Lee, K. Ottenberg, and M. Nölle., "Analysis and solutions of the three point perspective pose estimation problem," in *In Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1991.

[31] G. R.Jones, D. Lee, N. S.Holliman, and D. Ezra, "Controlling perceived depth in stereoscopic images," in *Proc. SPIE Stereoscopic Displays and Virtual Reality Systems VIII*, vol. 4297, Jun. 2001, pp. 42–53.

[32] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 313–318, 2003.

[33] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Trans. Graph.*, vol. 26, no. 3, p. 10, 2007.

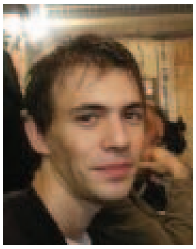
[34] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," *Proceedings of SIGGRAPH 2001*, pp. 341–346, August 2001.

[35] K. Tuite, N. Snavely, D. yu Hsiao, N. Tabing, and Z. Popovic, "Photocity: training experts at large-scale image acquisition through a competitive game," in *CHI*, 2011, pp. 1383–1392.

Chenxi Zhang received the B.S. degree from the Information Engineering, Zhejiang University, China in 2007. He is currently a Ph.D. candidate in the Computer Science Department at the University of Kentucky. His research interests include computer vision, computer graphics and image processing.

Jizhou Gao received the B.S. degree from the Computer Science Department, Zhejiang University, China in 2006. He is currently a Ph.D. candidate in the Computer Science Department at the University of Kentucky. His research interests include computer vision, computer graphics and data mining.

Oliver Wang is an Associate Research Scientist with Disney Research Zurich. He received his PhD in Computer Science in 2010 from the University of California, Santa Cruz in the area of computer graphics and image processing. He has applied his research expertise in numerous fields in both academia and industry, including doing a post-doctoral position with Disney Research and research internships at HP Labs, Industrial Light and Magic, and the Max Planck Institute for Informatik.



Pierre Georgel is a principal engineer at Dekko Inc. He received his Ph.D. in computer science in 2011 from the Technical University Munich, Germany. He published over 20 peer reviewed papers in conferences and journals.

Ruigang Yang received the MS degree in Computer Science from Columbia University in 1998 and the PhD degree in Computer Science from the University of North Carolina, Chapel Hill, in 2003. He is an Associate Professor in the Computer Science Department, University of Kentucky. His research interests include computer vision, computer graphics, and multimedia. He is a member of the IEEE Computer Society and the ACM. He is a recipient of NSF Faculty Early Career Development (CAREER) Program Award in 2004. He is an associate editor of IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI).

James Davis is an Associate Professor in Computer Science at University of California Santa Cruz. His research areas include computer graphics, visual sensing systems, computational photography, human computation, and ICTD. This work has resulted in over 80 peer-reviewed publications, patents, and invited talks, an NSF CAREER award, and best paper awards at ICRA 2003 and ICCV 2009. He was the founding director of the Center for Entrepreneurship (C4E) at University of California Santa Cruz. His teaching has twice been awarded for innovative style, including a course on the importance of technology to social entrepreneurship. He is on the advisory boards of several for-profit and non-profit organizations. He received his PhD from Stanford University in 2002, and was a senior research scientist at Honda Research Institute 2002-2004.

Jan-Michael Frahm is an Assistant Professor at University of North Carolina at Chapel Hill where he heads the 3D computer vision group and he is director of Computer Vision at RENCI (RENaissance Computing Institute). He received his Ph.D in computer vision in 2005 from the Christian-Albrechts University of Kiel, Germany. Dr. Frahm has worked on a variety of topics in geometric computer vision, the interface between geometric computer vision and recognition, realtime computer vision and active computer vision. Dr. Frahm is editor in chief for the Elsevier journal on image and vision computing.

Marc Pollefeys is a full professor in the Dept.of Computer Science of ETH Zurich since 2007. He also remains associated with the Dept. of Computer Science of UNC at Chapel Hill where he started as an assistant professor in 2002. He holds a Ph.D. (1999) from the KULeuven. His main area of research is geometric computer vision and his aim is to develop flexible approaches to capture visual representations of real world objects, scenes and events. Prof. Pollefeys received a Marr prize, NSF CAREER award, Packard Fellowship and ERC grant. He is the author of more than 170 peer-reviewed publications. He is the General Chair for ECCV2014 and was a Program Co-Chair for CVPR2009. He is on the Editorial Board of IJCV, an associate editor for the IEEE PAMI and is a Fellow of the IEEE.