DiffusionBAS: Estimating Camera External Orientation Through Diffusion Process

V.V. Kniaz[1,2], T.N. Skrypitsyna [3], V.A. Knyaz[1,2], S.Yu. Zheltov[2]

[1] Moscow Institute of Physics and Technology (MIPT), Moscow, Russia - (kniaz.va, kniaz.vv)@mipt.ru
[2] State Research Institute of Aviation Systems (GosNIIAS), Moscow, Russia - zhl@gosniias.ru
[3] Moscow State University of Geodesy and Cartography (MIIGAiK), Moscow, Russia

Keywords: cultural heritage, neural networks, diffusive network model, generative adversarial learning.

Abstract

Violent forces such as earthquakes or human interaction can damage or demolish objects of cultural heritage. Many architectural masterpieces have survived only in a few photos or drawings. Moreover, often interior decorations of buildings such as stucco or paintings are destroyed by fire. Therefore, an automatic 2D-to-3D reconstruction that can assists architecture historians in process of restoration of original 3D shape of a lost site of culture heritage is required. An automatic in-paint method can assists restoration of partially destroyed stucco and paintings. We present an end-to-end framework that receives a single image of an object and predicts its vector of the exterior orientation parameters. The main objective of a present work is reconstruction of 3D model of a partially destroyed 3D object and it 3D in-painting. As the initial and prerequisite phase of this framework we propose a new generative model for estimation of the exterior orientation parameters for a given input image using a Diffusion model (`DiffusionBAS`).

1. Introduction

Lost in the field. It is hard to imagine a more striking picture of abandonment and depressure. Still, an existing physical part of destroyed 3D object holds multiple traces of the original appearance. The task of reconstruction of completely destroyed objects of cultural heritage is much more challenging. Indeed, a method capable of reconstruction a partially destroyed building from a single photo must solve two non-determined tasks simultaneously:

1. Reconstruction of existing complete destroyed 3D building from a single view;
2. Reconstruction of the original 3D appearance of a destroyed building.

While single-photo 3D reconstruction had become an established technology in recent time, generation of 3D model of original shape of partially destroyed building remains challenging. This task is also commonly called 3D shape inpainting. Recently a NeRFiller model was proposed. NeRFiller model leverages two off-shelf available models to solve a challenging task of forecasting the appearance for lost parts of an object basing on its photograph.

It should be noted, that authentic recovering of the information about 3D scene is possible if accurate estimation of camera exterior orientation is performed. Traditional photogrammetric methods of determining of camera exterior orientation use a set of 3D points with known 3D coordinates. These points serve as reference information for the estimation procedure.

With the advances in developing data-driven methods new approaches based on machine learning have been proposed. This study addresses the problem of of blind external orientation parameters estimation from a single input image that is needed for accurate and holistic recovering the appearance of partly lost objects of cultural heritage.

The 2D and 3D in-painting techniques created the background for virtual reconstruction of the lost or damaged parts of images (scenes). Firstly, the NeRF (Mildenhall et al., 2020) neural network model is trained using available observations of a given 3D object. After that missing regions are masked by an operator. After that a diffusion model is used to perform a 2D inpainting of masked regions. After that the NeRF model is trained using the updated dataset.

We aim developing a new `DiffusionBAS` model capable to predict external orientation parameters for given input image. The `DiffusionBAS` solves the problem of bundle adjustment by machine learning methods. We hypothesize, that there is a correlation between a reference image in a multi-view stereo set and its exterior orientation, that can be learned by neural diffusion model. Specifically, we train an autoencoder model to encode the input image and its exterior orientation parameters into a latent code Z and to reconstruct it back to exterior orientation parameters for given image.

After that, we train an image encoder that provides a mapping from the input image in image set $A \in I^i$ to a latent code Z_i that encodes the exterior orientation vectors for the image set A. Using the latent code Z_i and the matrix decoder D, we reconstruct the the vector of exterior orientation parameters for input image.

Our key idea is to predict the vector of exterior orientation parameters from input image. Specifically we use a Stable diffusion model (Rombach et al., 2022) as a starting point for our `DiffusionBAS` model. We modify Stable Diffusion architecture to encode the input image and its exterior orientation parameters into a latent code and reconstruct it back to exterior orientation parameters for given image.

The main contributions of the study are the following:

- the framework for estimation of the exterior orientation of the single image based on a diffusion neural network model
- creation of the dataset for the proposed framework training and evaluation
- evaluation of the proposed framework on created dataset and baselines

2. Related work

Reconstruction of a complete 3D model from a limited or corrupted information is a challenging problem that has been analyzed by a scientific community for a long time. First approaches leveraged a combination of classical photogrammetry and analytic approximation (Mizginov and Kniaz, 2019, Kniaz et al., 2019). Photogrammetry reconstruction provided a detailed 3D model of the survived elements of a partially destroyed 3D object. The structure of the original (undestroyed) object was decomposed into a number of 3D primitives. Parameters of the primitives were optimized using linear least squares with an objective of minimizing surface distance between the surface of a primitive and fragment of a surface from photogrammetric reconstruction. More than ten years ago a rapid rise of neural networks had begun. Models was trained using a deep learning framework with an increased GPU performance and revolutionary architecture with convolution layers.

Deep learning models had dramatically changed the landscape of modern photogrammetry and 3D vision. In the field of photogrammetry deep learning had drastically improved the quality of feature point detection and matching (Yi et al., 2016a, Ono et al., 2018, Li et al., 2020). Such an improvement allowed to further densify the resulting point cloud. In the field of 3D vision some qualitatively new approaches had appeared.

Firstly, new models for depth map estimation from a single photo or a stereo pair (Zheng et al., 2018, Isola et al., 2017) have been proposed and evaluated in single object and arbitrary scene statements.

Secondly, a large number of neural network models for six degree of freedom camera orientation have been developed (Kendall et al., 2015) that can predict camera orientation with the accuracy suitable for qualitative task of machine vision.

Finally, new neural network models were designed for a single photo 3D model reconstruction. Unlike photogrammetric algorithms that required multiple views for generation of a complete 3D model, neural models allowed generation of an all-around 3D model from a single photo (Xie et al., 2019, Kniaz et al., 2020, Wu et al., 2017, Knyaz, 2020). Still reconstruction or 3D inpainting was challenging task for a regular volumetric convolutional neural networks.

Only the invention of diffusion models allowed effective restoration of missing data in 3D space. The NeRFiller model (Weber et al., 2023) solved this challenging task using a combination of a neural radiance field model (NeRF) (Mildenhall et al., 2020) and a diffusion model (Rombach et al., 2022).

The problem is solved iteratively. Firstly an approximate NeRF model is trained using available images of a 3D object. After that novel views of object a rendered using the NeRF approximation. After that diffusion model is used to reconstruct the original object appearance in synthetic novel views. Finally a new NeRF approximation is trained using original and new images.

Still NerFiller model requires multiple images to reconstruct the missing parts of 3D object. Our aim is to develop a new `DiffusionBAS` model capable of simultaneous 3D reconstruction and 3D in-painting of missing parts. At the current phase of the study, we address the problem of robust and accurate exterior orientation in the frame of whole recovering process. We aim developing a mapping $G : I \rightarrow Z$, where I is an arbitrary image of an object of interest, Z is a latent code encoding the vector of exterior orientation parameters.

Estimation of orientation parameters is one of the key elements of accurate 3D reconstruction of a scene using photogrammetric methods. In recent decades, various methods have been proposed to solve this problem, which vary depending on the available technical means and current image acquisition conditions.

These techniques usually exploit information about spatial coordinates of several reference scene points to estimate the parameters of image acquisition model, considering the images of these points as observations. The identification of the reference points in images firstly was performed manually by an operator, and later, with advances in image processing methods, by automatic detection algorithms. After establishing the correspondence between scene 3D points and their images, the parameters of imaging model are determined by bundle adjustment procedure based on least mean squares estimation.

The development of robust feature descriptors allowed to match corresponding features in different images, thus provided the basis for solving the task of ”structure-from-motion” – finding the camera orientation and reconstructing the 3D model from a set of images with unknown orientation.

The problem of reliable estimation of the orientation parameters attracted attention of the photogrammetric and computer vision scientific society from the first steps of image based 3D reconstruction (Hartley and Zisserman, 2004, Ozyesil et al., 2017, Knyaz and Zheltov, 2017, Kniaz et al., 2022).

The estimation is typically based on detection of some object (or a scene) points, that can be used as the reference data for available observations. Several robust and accurate descriptors such as Speed-Up Robust Features (SURF) or Scale Invariant Feature Transform (SIFT) (Bay et al., 2006, Lowe, 2004) have been developed, significantly improving the level of automation in correspondence problem solution. With the era of deep learning a set of neural network models were developed to solve the key-points detection problem (DeTone et al., 2018, Yi et al., 2016b).

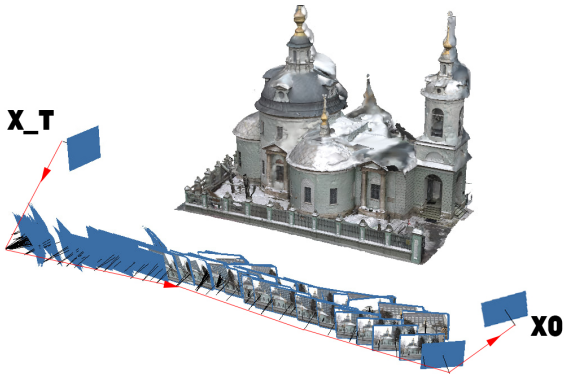

Figure 1. Diffusion process in 3D

The invention of the diffusion neural networks allows to apply such approach for the various task of computer vision and

3D reconstruction (Figure 1), such as image generation (Song and Ermon, 2019, Ho et al., 2020) or 3D point cloud generation (Luo and Hu, 2021, Lyu et al., 2021, Melas-Kyriazi et al., 2023) . The diffusion neural networks take their origin from non-equilibrium thermodynamics (Sohl-Dickstein et al., 2015). The learn to represent the distribution of analysed data by a Markov Chain, performing iterative adding noise to the initial data.

They demonstrate high performance in generating diverse high-quality samples, that makes this kind of neural network models to be an encouraging mean for direct exterior orientation estimation.

3. Materials and Method

In this paper, we propose a new generative model for implicit Bundle adjustment in the task of exterior parameters estimation. Diffusion neural network model (`DiffusionBAS`) is proposed as alternative to standard Bundle adjustment procedure.

The problem of the exterior orientation for a set of images of the same scene can be formulated as follows. Let $I^i \in \mathbb{R}^{3\times H\times W}$ be a set of the images of the same scene. It is required to find a set of corresponding vectors of exterior orientation parameters v_{eo} for this image set.

Exterior orientation maps a 3D point $\mathbf{p}_w \in \mathbb{R}^3$ from world coordinates to a 3D point $\mathbf{p}_c \in \mathbb{R}^3 = g^i(\mathbf{p}_w)$ in camera coordinates

3.1 Bundle adjustment problem

Bundle adjustment problem consists in estimating a vector of unknown parameters of the mathematical model used for describing the process of image formation on the sensor's matrix.

The basic imaging model used in photogrammetry is co-linearity equations. They formulate the fact of belonging to the same ray the following points: the point of the scene S, the center of the projection of C and the projection s of the point S in the image plane:

$$\mathbf{x}_s - \mathbf{x}_p = -\lambda \mathbf{R}\cdot(\mathbf{X}_S - \mathbf{X}_C) \tag{1}$$

The next notations are used in Equation 1:

$\mathbf{X}_C = (X_C, Y_C, Z_C)^T$ – coordinates of the center of the projection,
$\mathbf{X}_S = (X_S, Y_S, Z_S)^T$ – coordinates of the scene point S ,
$\mathbf{x}_s = (x_s, y_s, -c)^T$ – the corresponding coordinates of the scene point S in the image,
$\mathbf{R}$ – coordinate system rotation matrix,
$\mathbf{x}_p$ – coordinates of the principal point,
λ – scale factor.

The coordinates of point S and of the center of projection C are defined in object coordinate system $OXYZ$ (Figure 2), coordinates of image point $\mathbf{x}_s$ are defined in image coordinate system $Cxyz$. The transition from object coordinate system to image coordinate system is determined by the matrix $\mathbf{R}$:

$$\mathbf{R} = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33}, \end{pmatrix}, \tag{2}$$

and the elements r_{ij} of the matrix $\mathbf{R}$ are defined by Euler's rotation angles α, ω, κ.

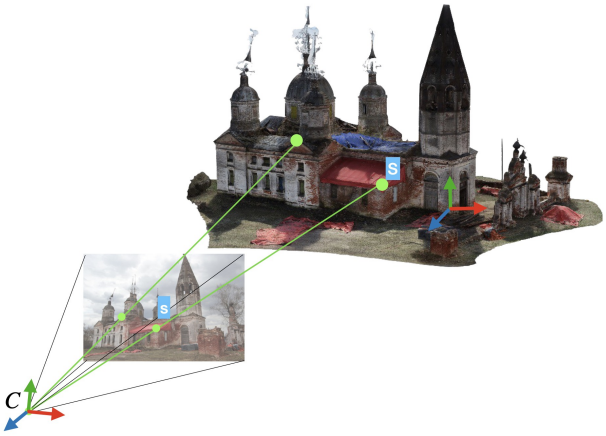

Figure 2. Co-linearity condition

Equations 1 can be written in form:

$$x_s - x_p = -c\frac{r_{11}(X_S - X_C) + r_{12}(Y_S - Y_C) + r_{13}(Z_S - Z_C)}{r_{31}(X_S - X_C) + r_{32}(Y_S - Y_C) + r_{33}(Z_S - Z_C)}$$
$$y_s - y_p = -c\frac{r_{21}(X_S - X_C) + r_{22}(Y_S - Y_C) + r_{23}(Z_S - Z_C)}{r_{31}(X_S - X_C) + r_{32}(Y_S - Y_C) + r_{33}(Z_S - Z_C)}$$

The task of exterior orientation is to find the values of vector of exterior orientation parameters $v^{eo} = (X_C, Y_C, Z_C, \alpha, \omega, \kappa)^T$ for given image. With bundle adjustment technique this problem is solved as the task of estimating unknown parameters basing on observations. The aim of the bundle adjustment procedure is to minimize the errors in calculating of the 3D coordinates when applying the imaging model (Equation 1).

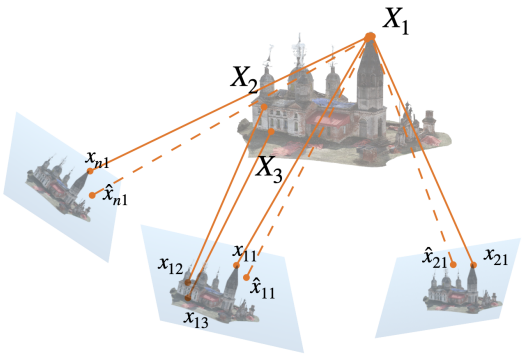

Figure 3. Data for Bundle Adjustment procedure.
x_{ij} are the image coordinates (observations of X_j) in i^{th} image, X_j are the spatial coordinates of object points, $\hat{x}_{ij}(v_i^{eo})$ are the re-projected x_{ij} points.

This criterium can be written in different forms, depending on available data. In the case, when reference 3D coordinates $\{X_j, j = 1, \ldots, m\}$ are given, the L_{3D} metric is to be min-

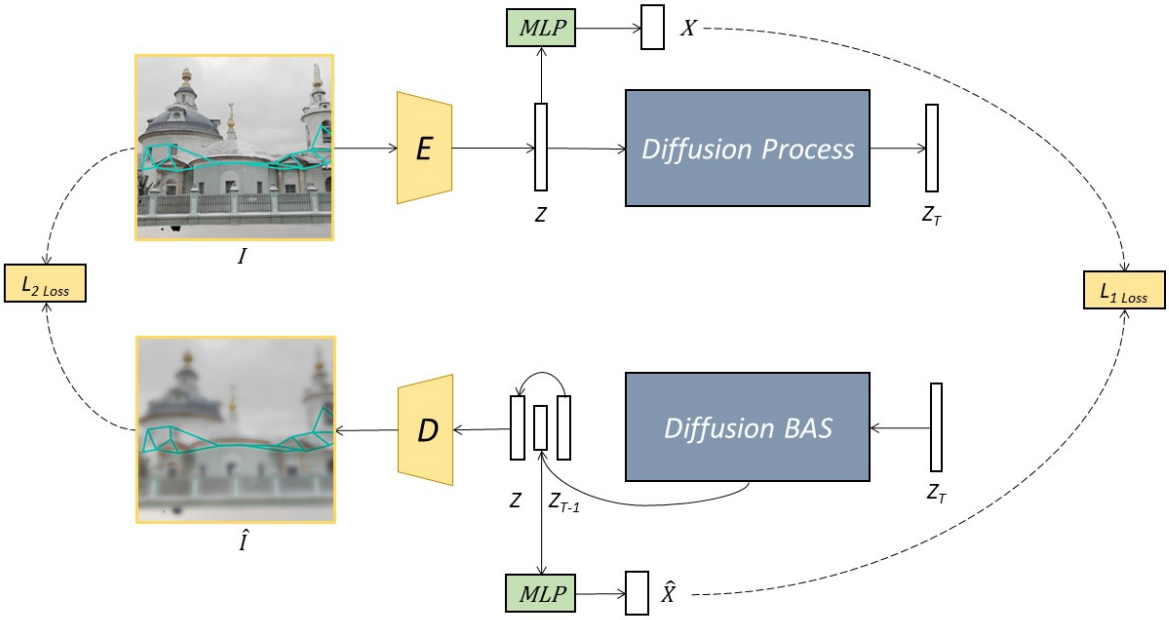

Figure 4. Diffusion BAS model architecture

imized:

$$L_{3D} = \sum_{i=1}^{n}\sum_{j=1}^{m}(X_j - \hat{X}_{i,j}(v_i^{eo}, x_{ij}))^2 \tag{3}$$

If the information about 3D coordinates of the scene is not available (as for structure from motion or simultaneous localization and mapping problems), the criterium can be written as re-projection error of some points of the scene

$$L_{rp} = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \hat{x}_{i,j}(v_i^{eo}, x_{ij}))^2 \tag{4}$$

where x_{ij} are the images coordinates of object point X_j (observations), representing the j^th 3D object point X_j in the i^th image. $\hat{x}_{i,j}(v_i^{eo}, x_{ij})$ is defined by the non-linear transformation of Equation 1.

3.2 Diffusion BAS framework overview

Recently proposed diffusion neural networks (Ho et al., 2020, Sohl-Dickstein et al., 2015, Song and Ermon, 2019) can be considered as another way of the iterative estimation of exterior orientation by bundle adjustment. They firstly learn the possible distribution of the disturbed data by sequentially adding noise to input data. At the inference phase the trained diffusion neural network predicts the undisturbed data from given input sample by reverse deniosing process.

We hypothesise that such approach can improve the bundle adjustment performance for arbitrary set of scene images, overcoming the problem of poor convergence of bundle adjustment procedure with ”bad” initial conditions.

The architecture of the proposed `DiffusionBAS` model is shown in Figure 4. We modify Stable Diffusion (Rombach et al., 2022) architecture to encode the input image and its exterior orientation parameters into a latent code and reconstruct it back to exterior orientation parameters for given image.

3.3 Camera External Orientation Estimation through Reverse Diffusion

We consider an estimation of camera external orientation parameters as a reverse diffusion process. We summarize our approach in the Algorithm 1.

Algorithm 1: Reverse Diffusion

Data: A set of images with known external orientation $\mathcal{I} = \{I_0, I_1, \dots I_K\}$, a set of known ground control points $\mathcal{P}_{GCP}$, an input image I for which an external orientation should be found.

Result: X_0 = `DiffusionBAS`(I)

$D_{\min} \leftarrow \infty$;

while $j < K$ **do**
  $D_j \leftarrow FID(I_j, I)$;
  if $D_j < D_{\min}$ **then**
    $D_{\min} \leftarrow D_j$;
    $j_{\min} \leftarrow j$;

$I_{ref} \leftarrow I_{j_{\min}}$;
$X_t \leftarrow X_{j_{\min}}$;
$t \leftarrow T$ **while** $t > 0$ **do**
  $I'_t \leftarrow$ `Plot`$(\mathcal{P}_{GCP}, I_t)$;
  $\Delta X \leftarrow$ `DiffusionBAS`(I'_t);
  $X_{t-1} \leftarrow X_t + \Delta X$;

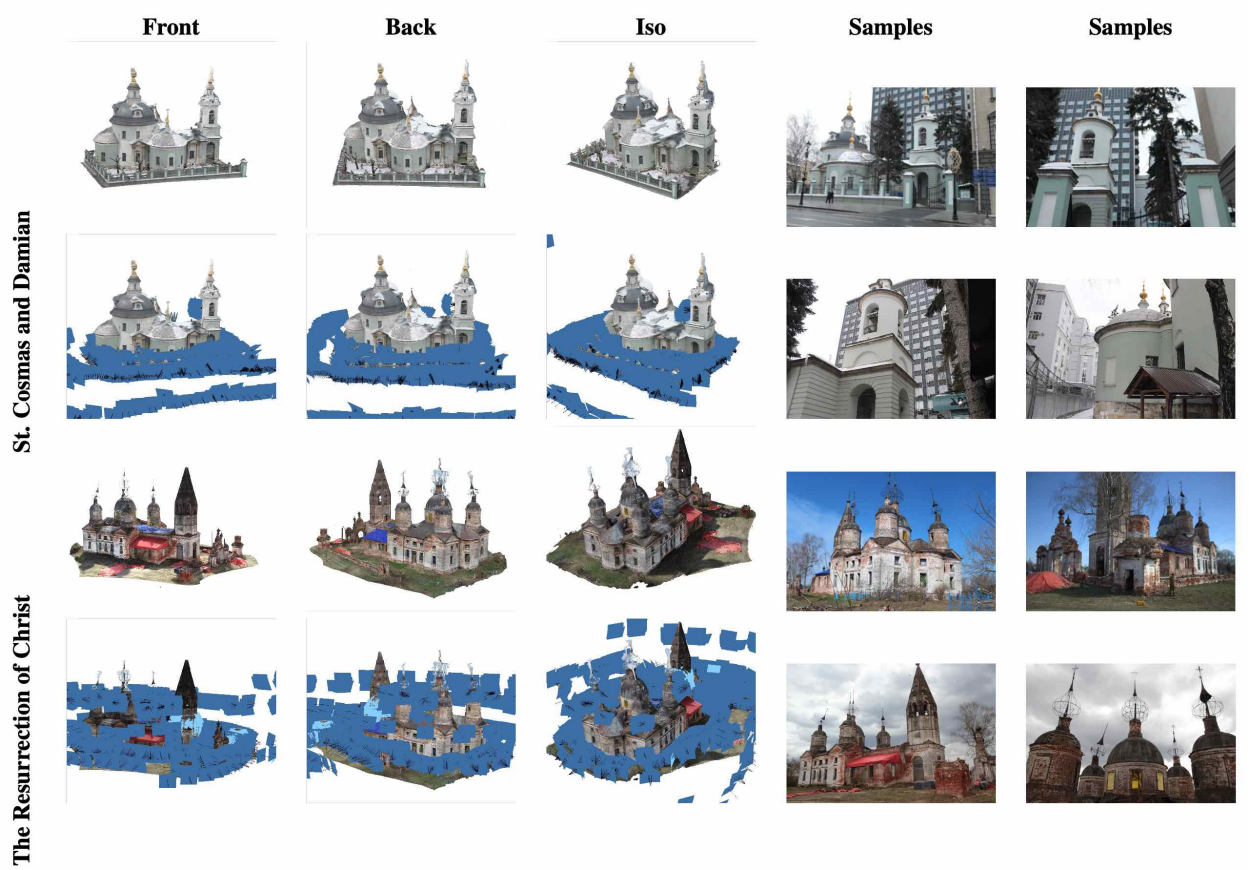

Figure 5. Examples of annotated images in our *Heritage3D* dataset.

3.4 Dataset generation

Addressing the problem of recovering the appearance of cultural heritage objects, we developed a new *Heritage3D* dataset containing samples with 2D and 3D representation of partially destroyed objects of cultural heritage located in central Russia.

Each sample includes the following data: an image of an object and the ground truth camera external orientation parameters in the object coordinate system. Also we provide ground truth 3D low polygonal models of objects generate by manual processing of a rough 3D model produced by a structure-from-motion pipeline.

Our dataset includes 4 objects of cultural heritage. For each object we include 50 real images, and 5k images generated using a NeRF trained from objects photos. Example images from our dataset are presented in Figure 5.

4. Evaluation Results

We evaluated our `DiffusionBAS` model and standard bundle adjustment procedure on our *Heritage3D* dataset in terms of mean square error between the estimated camera external orientation parameters and the ground truth ones. The results of the evaluation are presented in Table 1.

Table 1 shows, that the developed `DiffusionBAS` framework for exterior orientation by implicit bundle adjustment using diffusion-based machine learning outperforms the standard bundle adjustment technique and can compete with modern neural 6DOF pose estimation models.

Mean error of external parameter estimation		
Parameter	Diffusiion BA	Standard BA
X, m	0.0543	0.0621
Y, m	0.0371	0.0427
Z, m	0.0622	0.0918
α^o	0.0327	0.0511
ω^o	0.0271	0.0312
κ^o	0.0194	0.0302

Table 1. Results of `DiffusionBAS` evaluation.

5. Conclusion

We developed a new model for estimation of camera external orientation with known interior orientation parameters. Our `DiffusionBAS` model is capable of blind external orientation parameters estimation from a single input image. Our model leverage diffusion process. We trained an autoencoder model to encode the reference images and exterior orientations from the dataset into a latent code Z and then to reconstruct back the exterior orientation vector from the latent code Z for given input image.

We collected the new *Heritage3D* dataset containing 2D and 3D samples for partially destroyed objects of cultural heritage. We evaluated our model on our *Heritage3D* dataset. The evaluation showed the our model outperforms hand crafted methods and can compete with modern neural 6DOF pose estimation models.

Acknowledgements

The research was carried out at the expense of a grant from the Russian Science Foundation No. 24-21-00314, https://rscf.ru/project/24-21-00314/

References

Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: Speeded up robust features. A. Leonardis, H. Bischof, A. Pinz (eds), *Computer Vision – ECCV 2006*, Springer Berlin Heidelberg, Berlin, Heidelberg, 404–417.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. Self-supervised interest point detection and description'. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.*

Hartley, R., Zisserman, A., 2004. *Multiple View Geometry in Computer Vision*. Second edn, Cambridge University Press, ISBN: 0521540518.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840–6851.

Isola, P., Zhu, J. Y., Zhou, T., Efros, A. A., 2017. Image-to-image translation with conditional adversarial networks. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017-Janua, 5967–5976.

Kendall, A., Grimes, M., Cipolla, R., 2015. Convolutional networks for real-time 6-DOF camera relocalization. *CoRR*, abs/1505.07427. http://arxiv.org/abs/1505.07427.

Kniaz, V. V., Knyaz, V. A., Mizginov, V., Bordodymov, A., Moshkantsev, P., Barylnik, S., Novikov, D., 2022. IMAGE ORIENTATION BY EMBEDDING IN A GAN LATENT SPACE. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2/W2-2022, 149–155. https://isprs-archives.copernicus.org/articles/XLVIII-2-W2-2022/149/2022/.

Kniaz, V. V., Knyaz, V. A., Remondino, F., Bordodymov, A., Moshkantsev, P., 2020. Image-to-voxel model translation for 3d scene reconstruction and segmentation. A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (eds), *Computer Vision – ECCV 2020*, Springer International Publishing, Cham, 105–124.

Kniaz, V. V., Remondino, F., Knyaz, V. A., 2019. GENERATIVE ADVERSARIAL NETWORKS FOR SINGLE PHOTO 3D RECONSTRUCTION. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W9, 403–408. https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-2-W9/403/2019/.

Knyaz, V., 2020. Machine learning for scene 3D reconstruction using a single image. *Proc. SPIE 11353, Optics, Photonics and Digital Technologies for Imaging Applications VI*, 11353, 393 – 402. https://doi.org/10.1117/12.2556122.

Knyaz, V. A., Zheltov, S. Y., 2017. Accuracy evaluation of structure from motion surface 3D reconstruction. *Proc.SPIE*, 10332, 10332 - 10332 - 10. http://dx.doi.org/10.1117/12.2272021.

Li, J., Hu, Q., Ai, M., 2020. RIFT: Multi-Modal Image Matching Based on Radiation-Variation Insensitive Feature Transform. *IEEE Transactions on Image Processing*, 29, 3296-3310.

Lowe, D. G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), 91-110. https://doi.org/10.1023/B:VISI.0000029664.99615.94.

Luo, S., Hu, W., 2021. Diffusion Probabilistic Models for 3D Point Cloud Generation. *arXiv: 2103.01458*. https://arxiv.org/abs/2103.01458.

Lyu, Z., Kong, Z., Xu, X., Pan, L., Lin, D., 2021. A conditional point diffusion-refinement paradigm for 3d point cloud completion. *arXiv preprint arXiv:2112.03530*.

Melas-Kyriazi, L., Rupprecht, C., Vedaldi, A., 2023. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12923–12932.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *arXiv: 1805.09662*. https://arxiv.org/abs/2003.08934.

Mizginov, V. A., Kniaz, V. V., 2019. EVALUATING THE ACCURACY OF 3D OBJECT RECONSTRUCTION FROM THERMAL IMAGES. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W18, 129–134. https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-2-W18/129/2019/.

Ono, Y., Trulls, E., Fua, P., Yi, K. M., 2018. LF-Net: Learning Local Features from Images. *arXiv: 1805.09662*. https://arxiv.org/abs/1805.09662.

Ozyesil, O., Voroninski, V., Basri, R., Singer, A., 2017. A Survey of Structure from Motion. *arXiv: 1701.08493*. https://arxiv.org/abs/1701.08493.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-Resolution Image Synthesis with Latent Diffusion Models. *CoRR*, abs/2112.10752. https://arxiv.org/abs/2112.10752.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S., 2015. Deep unsupervised learning using nonequilibrium thermodynamics. *International Conference on Machine Learning*, PMLR, 2256–2265.

Song, Y., Ermon, S., 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Weber, E., Hołyński, A., Jampani, V., Saxena, S., Snavely, N., Kar, A., Kanazawa, A., 2023. NeRFiller: Completing Scenes via Generative 3D Inpainting. *arXiv: 2312.04560*. https://arxiv.org/abs/2312.04560.

Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, W. T., Tenenbaum, J. B., 2017. MarrNet: 3D Shape Reconstruction via 2.5D Sketches. *Advances In Neural Information Processing Systems*.

Xie, H., Yao, H., Sun, X., Zhou, S., Zhang, S., Tong, X., 2019. Pix2Vox: Context-aware 3D Reconstruction from Single and Multi-view Images. *CoRR*, abs/1901.11153. http://arxiv.org/abs/1901.11153.

Yi, K. M., Trulls, E., Lepetit, V., Fua, P., 2016a. LIFT: learned invariant feature transform. *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, 467–483.

Yi, K. M., Trulls, E., Lepetit, V., Fua, P., 2016b. LIFT: Learned Invariant Feature Transform. *arXiv: 1603.09114*. https://arxiv.org/abs/1603.09114.

Zheng, C., Cham, T., Cai, J., 2018. T2Net: Synthetic-to-Realistic Translation for Solving Single-Image Depth Estimation Tasks. *CoRR*, abs/1808.01454. http://arxiv.org/abs/1808.01454.