University of Missouri

Deep Realistic Novel View Generation for City-Scale Aerial Images

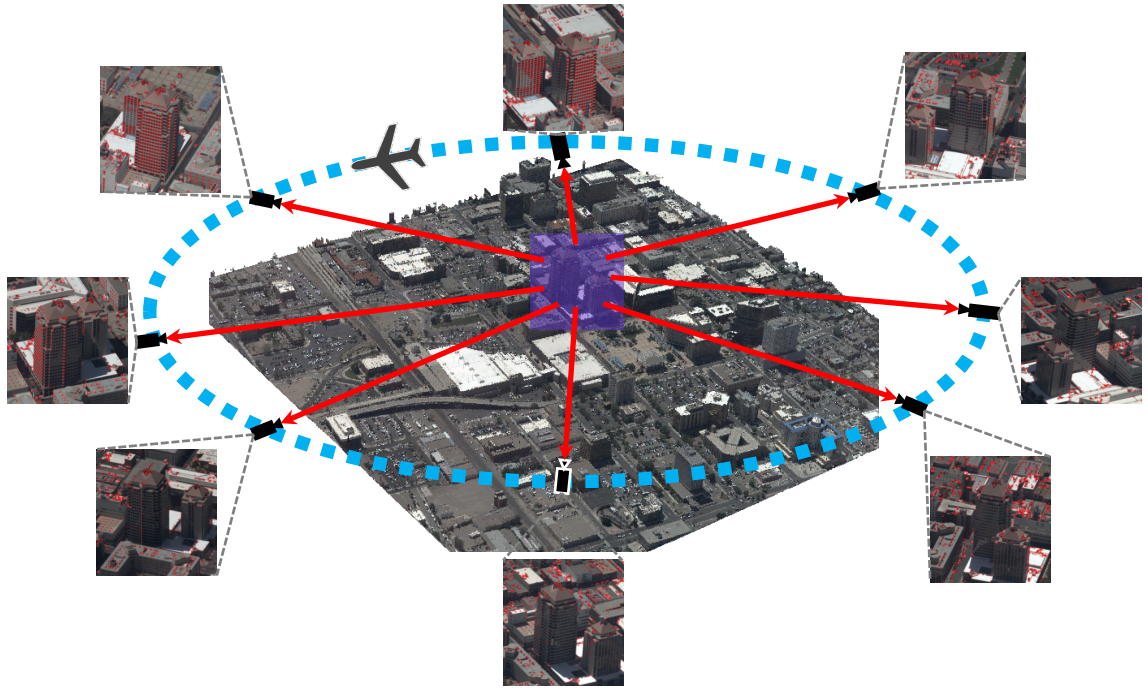
ICPR 2020

Koundinya Nouduri*, Ke Gao*, Joshua Fraser, Shizeng Yao, Hadi AliAkbarpour, Filiz Bunyak, Kannappan Palaniappan

University of Missouri-Columbia, MO, USA

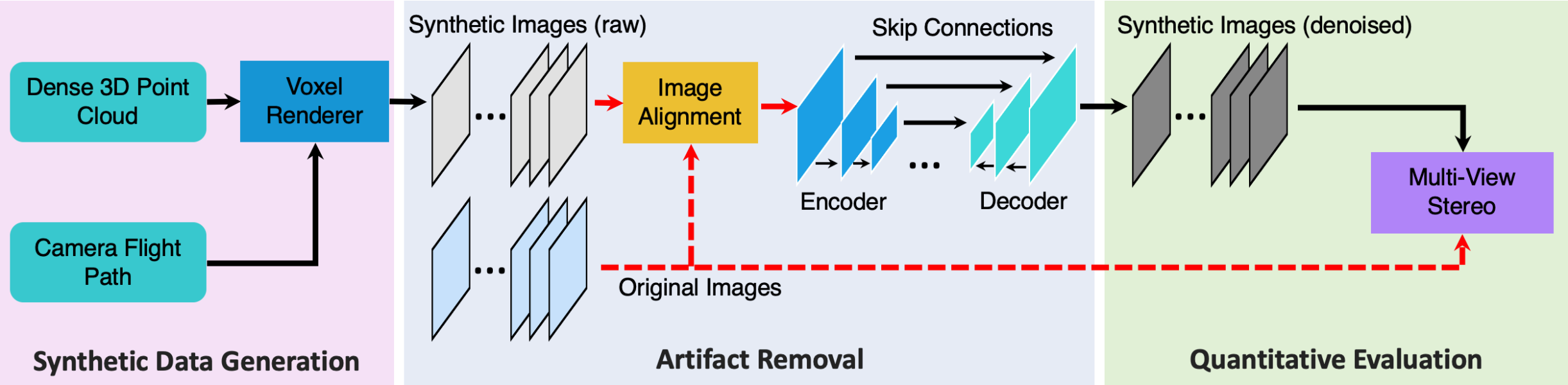
Motivation

- Performance of AI/ML based image analysis approaches depends significantly on the quantity and quality of the training data.
- Generation of annotated training data is often costly, time-consuming and laborious.
- Introduced a novel end-to-end framework for generation of large scale synthetic aerial image sequences with associated precise ground truth camera metadata
- Proposed a novel edge-augmented deep learning network with an explicit edgemap processing stream to remove image artifacts
- **Two main purposes**:
  1. Enable objective, quantitative evaluation of computer vision algorithms and methods such as feature detection, description, and matching or full computer vision pipelines such as 3D point cloud reconstruction.
  2. Provide large amounts of high-quality training data for deep learning guided computer vision methods

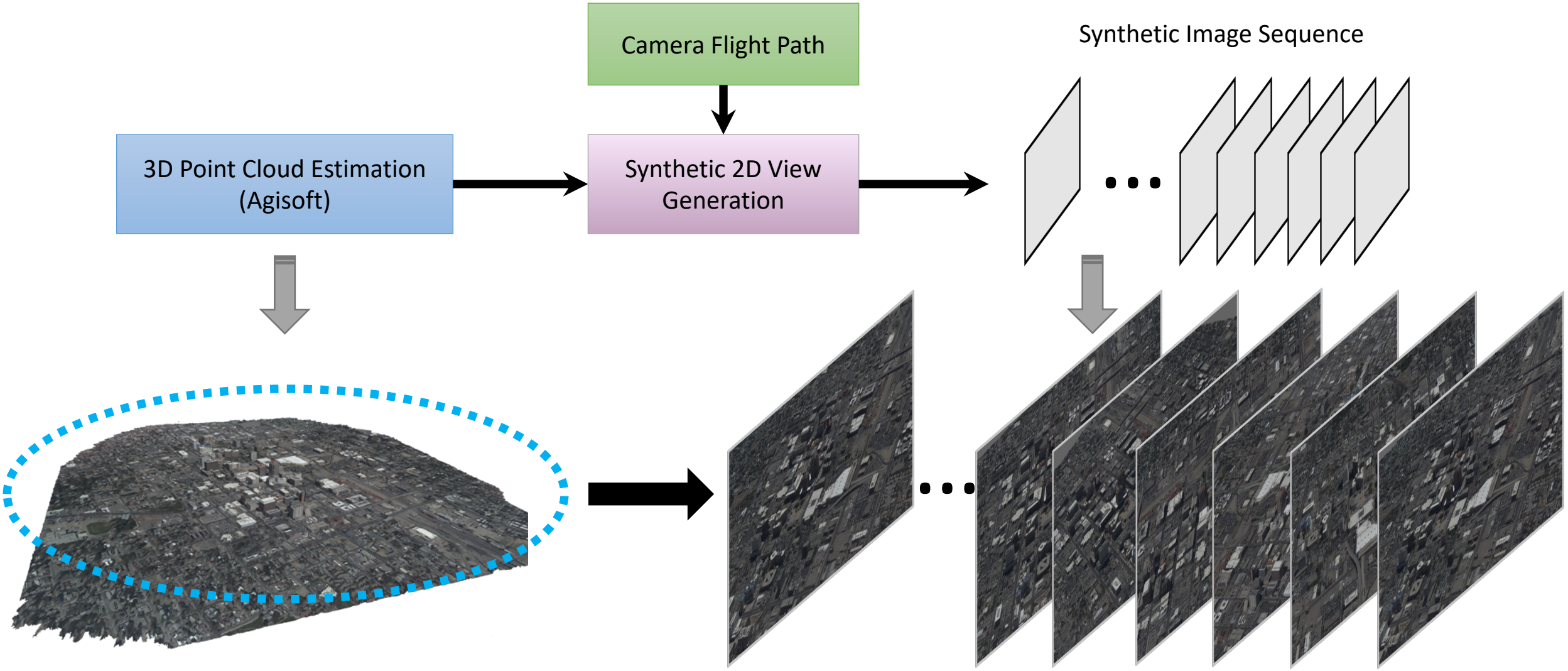


Proposed Pipeline

- Proposed pipeline for generating realistic novel views for city-scale aerial imagery from a dense 3D point cloud and a camera flight path
- **Black arrows:** data and processing flow that occur during **training and inference**
- **Red arrows:** data flow that occurs during **only training** such as image alignment between synthetic image and real image
- Original images not required at the point of inference
- Original images can be used as ground truth for quantitative evaluation for computer vision applications such as Multi-View Stereo

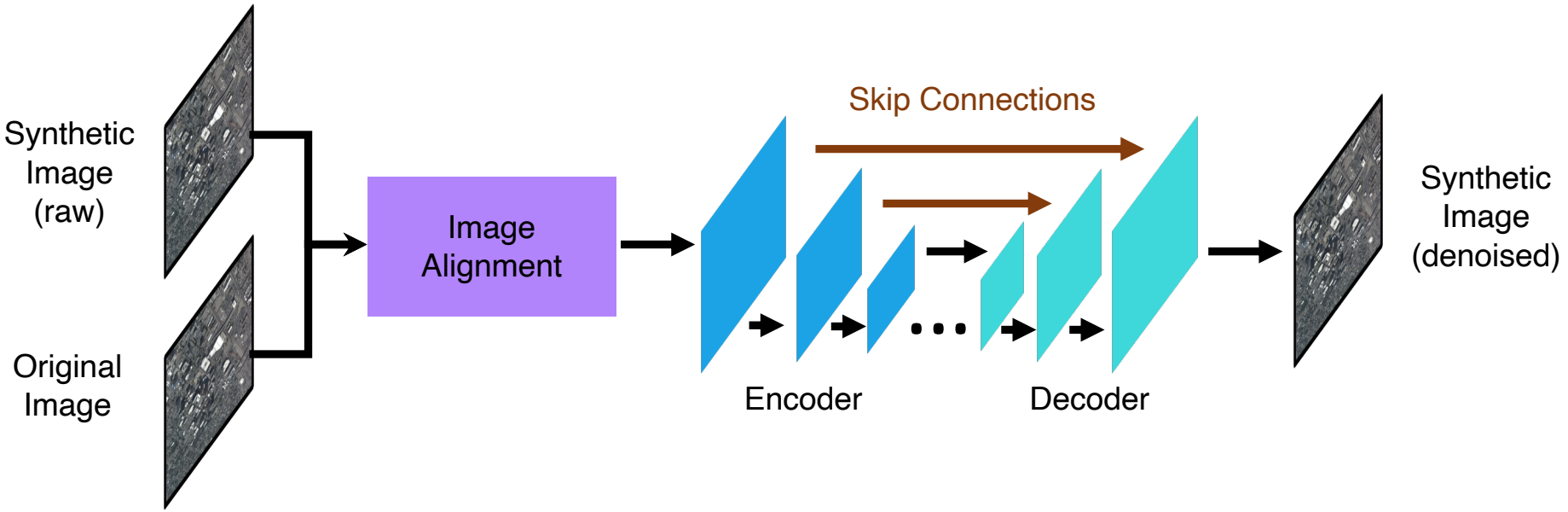


3D Voxel Renderer



Synthetic Image Artifact Removal

- Reconstruction of accurate dense 3D point clouds for large scale urban scenes is a challenging task with error prone results
- Developed a deep artifact removal network for 3D point cloud generated synthetic images
- Two large scale aerial image sequences of urban scene, Albuquerque (ABQ) and Los Angeles (LA), are used as ground truth for training and testing of the proposed artifact removal network
- Original aerial Image sequences provided by TransparentSky [1]



[1] http://transparentsky.net

Motivation for Image Artifact Removal

- Synthetic images generated from dense point clouds are prone to images artifacts caused by point cloud errors

Image patch size: 263x212

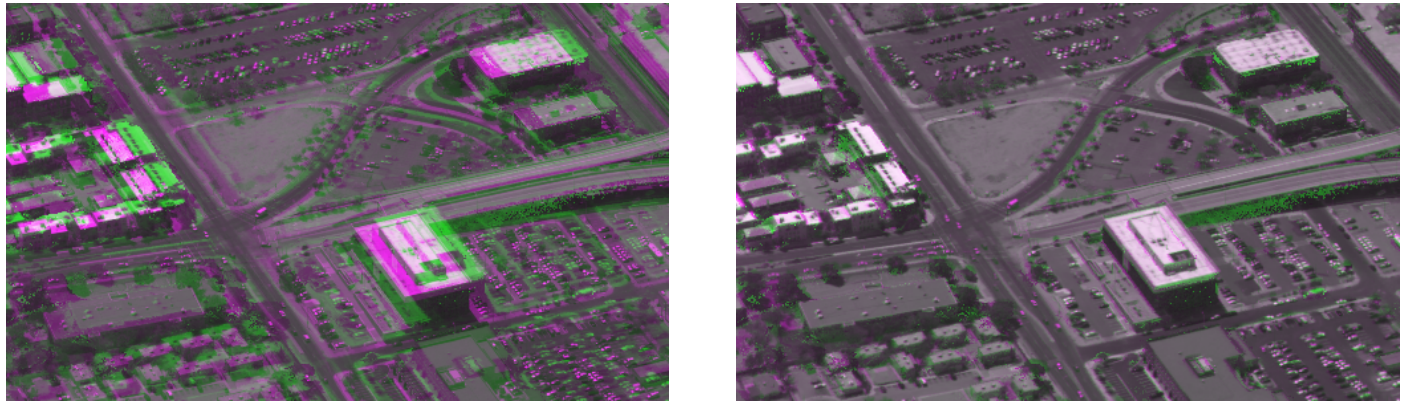
ABQ-215 Synthetic Data (image size: 1650 x 1100)

Training Data Processing

- Point cloud from Agisoft [1] manually aligned to the coordinate system of the original aerial data.
- Each raw synthetic image in the sequence is slightly shifted with respect to the corresponding original aerial image.
- If not corrected, the misalignment of the input and target image pairs can compromise the artifact removal performance during network training.

Registration of Original Real & Synthetic Images:

1. Perform feature matching (ORB features) between the original aerial image and the corresponding raw synthetic image.
2. Estimate the homography matrix using RANSAC
3. Warp the original images to the associated synthetic images prior to feeding them to the proposed network during training.



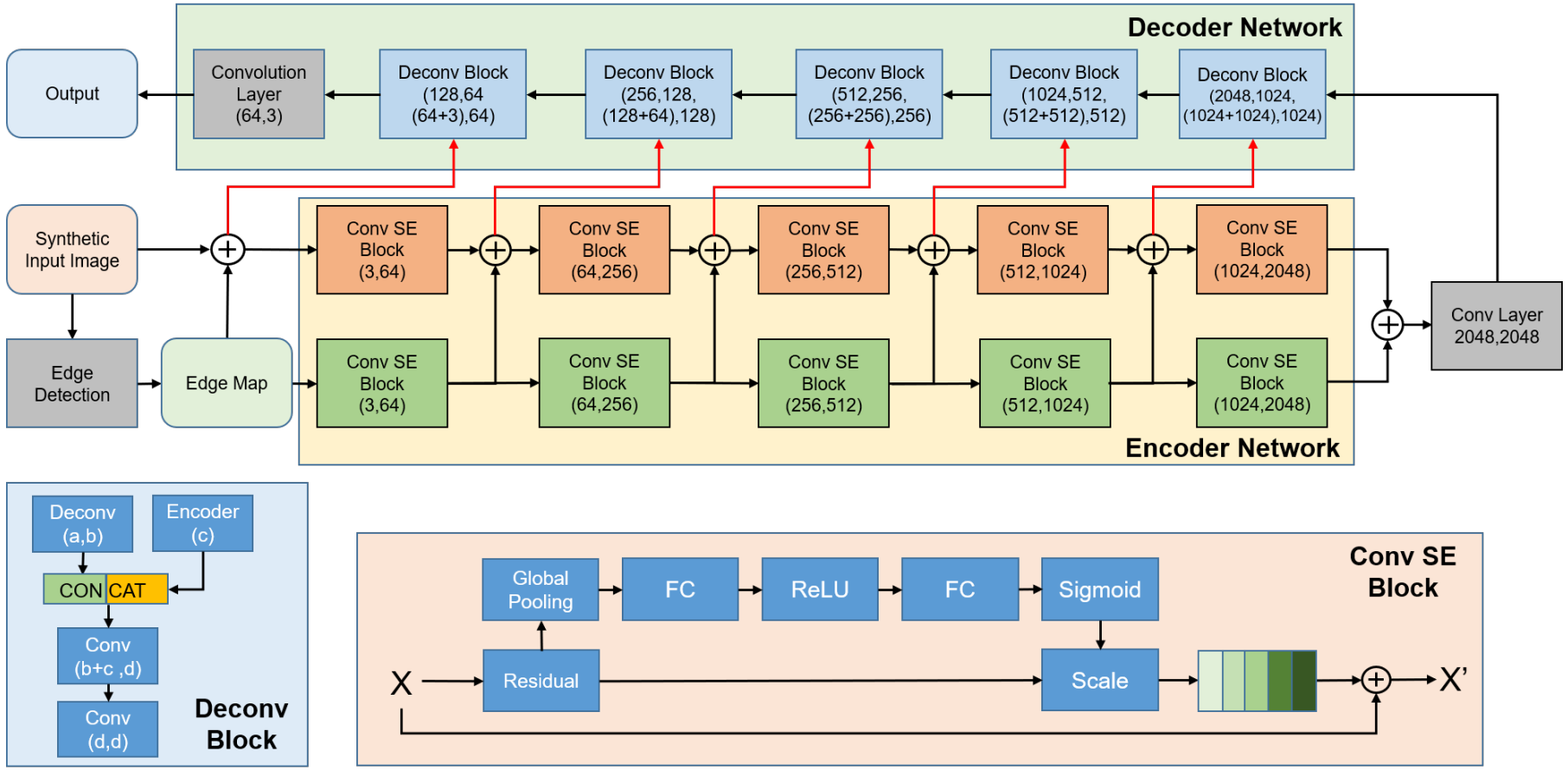
(a) (b)

Visualization of image alignment for data preprocessing. (a) original aerial image (magenta) vs. raw synthetic image (green). (b) aligned original aerial image (magenta) vs. raw synthetic image (green).

[1] https://www.agisoft.com/

Artifact Removal Network (ARNET-Edge)

- ARNET-Edge is our custom encoder-decoder architecture that learns to map raw synthetic images to realistic looking synthetic images.
- **Input:** three channel RGB images + grayscale canny edge map
- **Output:** denoised image

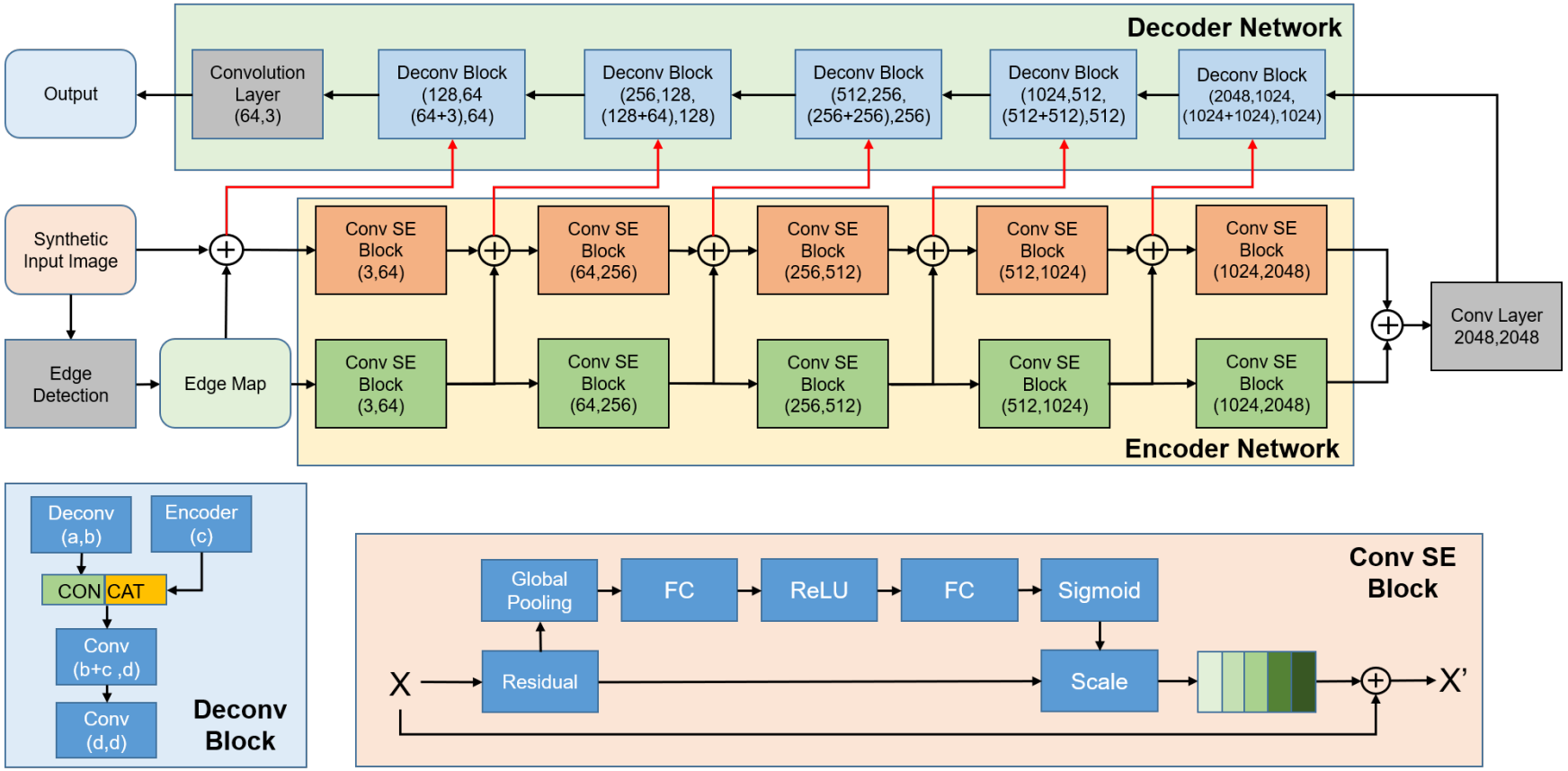


ARNET-Edge Architecture

Encoder: two parallel series of Squeeze and Excitation Resnet (SE-Resnet Blocks).

- **Stream-1:** extracts features from RGB input
- **Stream-2:** extracts features from edge-map
- Output of feature vectors after each convolution block of SE-Resnet module after summed to first stream.
- Encoder and decoder are connected by a 2048 output features.

Decoding block: deconvolution layer followed by two convolution layers attempting to reconstruct the original image from encoders features.



ARNET-Edge Loss Function

- Used combination of Mean Squared Error (MSE) and Structural Similarity Index Error (SSIM)
- SSIM shows the similarity between pair of images with 1 being most similar:

$$SSIM(gt, gi) = \frac{(2\mu_{gt}\mu_{gi} + C_1) + (2\sigma_{gtgi} + C_2)}{(\mu_{gt}^2 + \mu_{gi}^2 + C_1)(\sigma_{gt}^2 + \sigma_{gi}^2 + C_2)}$$

where gt is the ground truth image, gi is the image generated by the network, μ and σ^2 denote image mean and variance respectively, and C_1, C_2 are regularization constants.

- Structural similarity loss L_{ssim} and mean squared error loss L_{mse} are defined as:

$$L_{ssim}(gt, gi) = \frac{1}{N}\sum_{i=0}^{N}(1 - SSIM(gt, gi))$$

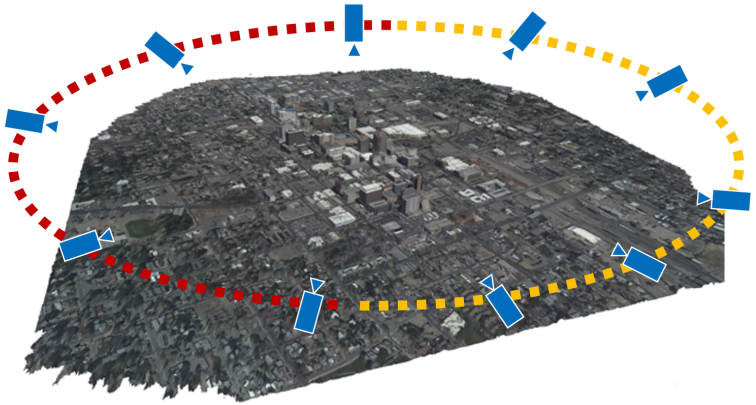
$$L_{mse}(gt, gi) = \frac{1}{N}\sum_{i=0}^{N}(gt - gi)^2$$

$$L_{total}(gt, gi) = L_{ssim}(gt, gi) + L_{mse}(gt, gi)$$

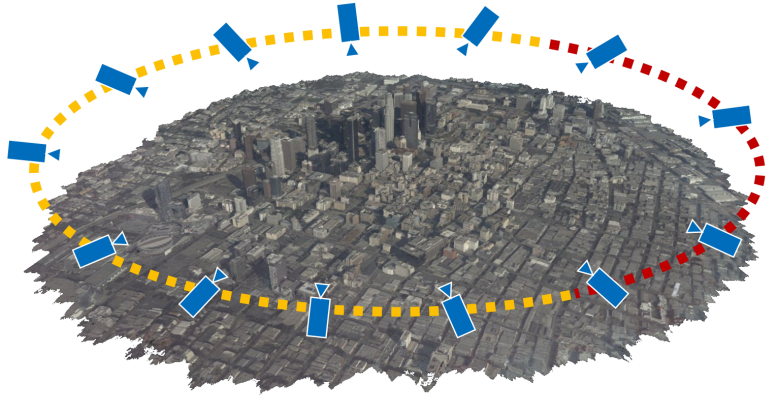
Dataset

- Two large scale aerial image sequences of urban scenes are used as ground truth for training and testing of the proposed artifact removal network
  - Albuquerque (ABQ) and
  - Los Angeles (LA)
- Data provided by Transparent Sky [1]
- High resolution RGB images were captured airborne and the flight trajectory was a full orbit around the downtown area of each city
- Synthetic aerial image sequences: generated using the voxel rendering process
- Dense 3D point clouds: produced by Agisoft

Dataset	# Images	# Training Images	# Test Images
ABQ	215	100	115
LA	351	100	251



ABQ



LA

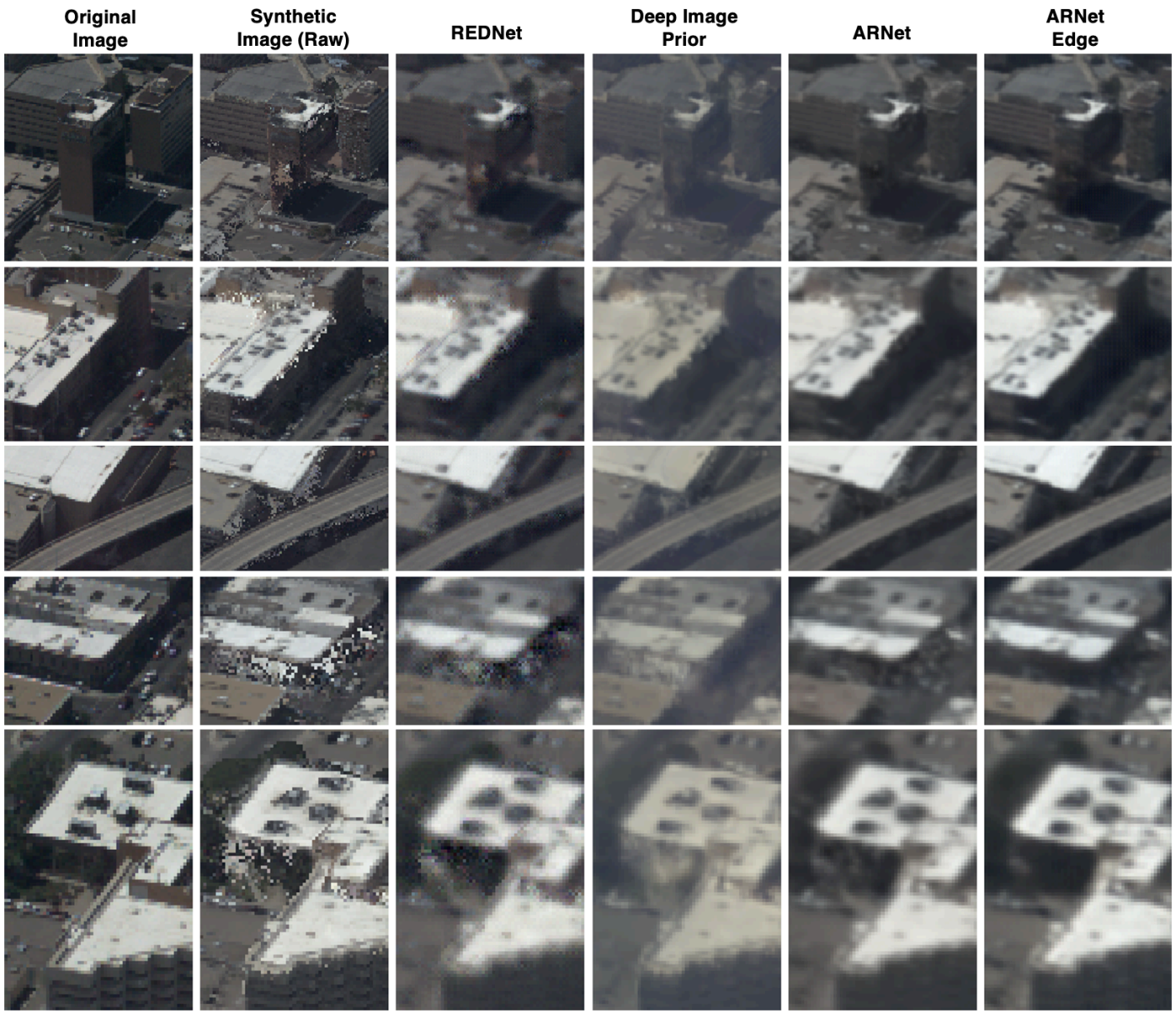
Visualization of camera flight path for ABQ and LA dataset. Red dots indicate training views and yellow dots indicate testing views. 100 views are used for training for both datasets

[1] http://transparentsky.net

Experimental Results

- We have developed ARNET a variation of our proposed network without edge-map feature extraction.

- We have compared our model with two other image restoration networks. It can been be visually seen that ARNET-Edge can recover building structures even under severe artifacts on edges and corners.

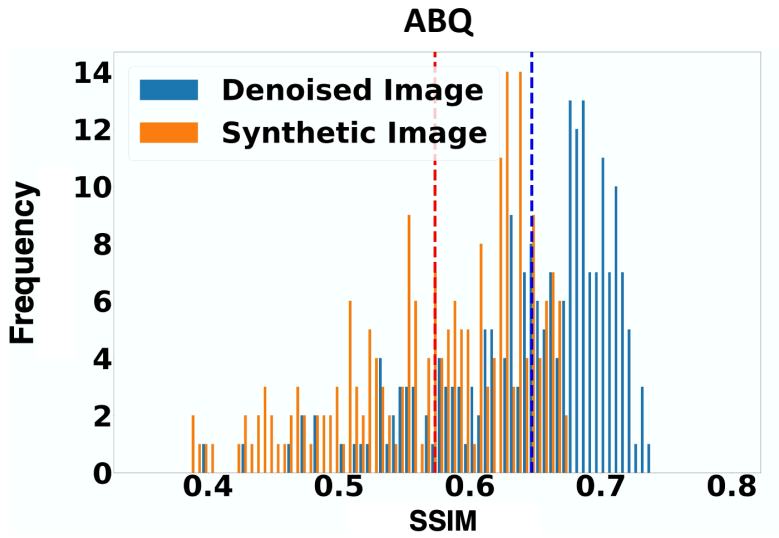


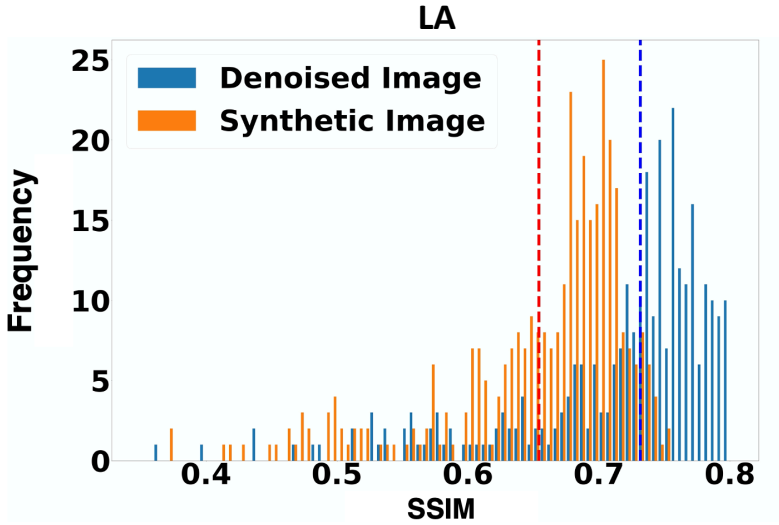
Experimental Results

Image Artifact Removal Performance Evaluation

- Compared the denoised image to original synthetic image in terms of Structural Similarity Index Measure (SSIM) and Peak Signal to Noise Ratio (PSNR)
- For each column, best results and second-best results are highlighted

Model	SSIM		PSNR	
	ABQ	LA	ABQ	LA
Raw Synthetic	0.573	0.651	20.65	23.53
REDNet [1]	0.629	0.686	22.29	24.35
Deep Image Prior [2]	0.619	0.710	21.12	25.21
ARNet	0.638	0.704	22.40	25.15
ARNet-Edge	0.665	0.721	23.21	25.89



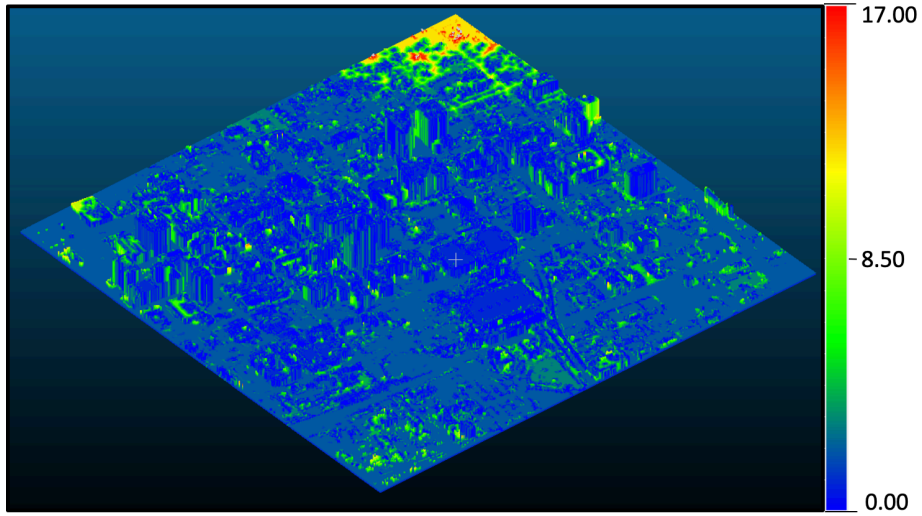


[1] X.-J. Mao, C. Shen, and Y.-B. Yang, “Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections,” in Proc. International Conference on Neural Information Processing Systems, 2016.
[2] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Deep image prior,” in IEEE Conference on Computer Vision and Pattern Recognition, 2018.

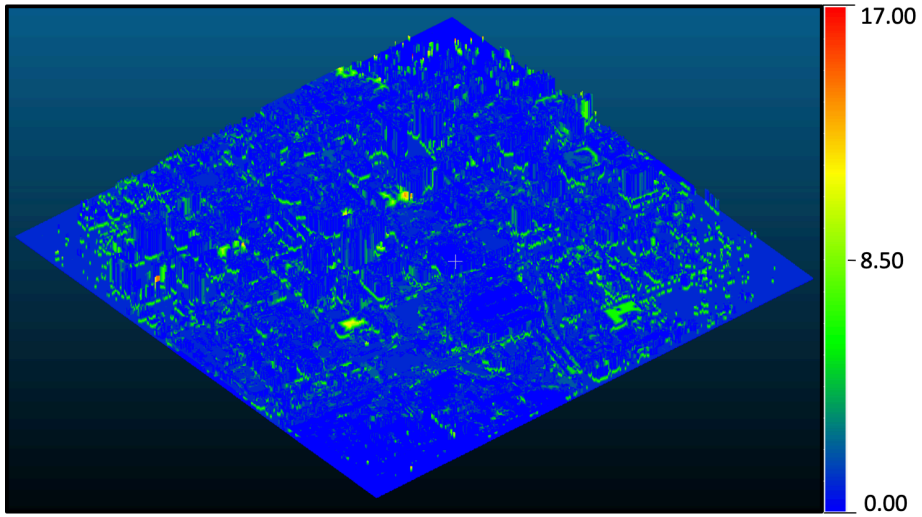
Experimental Results

Dense 3D Point Cloud Reconstruction

- Reconstructed a dense 3D point cloud from the denoised ABQ image sequence (PC_d) and from the raw ABQ image sequence (PC_r) using VB3D; point cloud generated from the original aerial image sequence as the ground truth (PC_0)
- Computed cloud-to-cloud distance of PC_r vs PC_0 and PC_d vs PC_0
- 24.70% of the points in PC_r have a distance less than 1 compared to PC_0; 69.24% of the points in PC_d have a distance less than 1 compared to PC_0



Cloud-to-cloud distance (PC_r vs ground truth PC_0)



Cloud-to-cloud distance (PC_d vs ground truth PC_0)

[1] S. Yao, H. AliAkbarpour, G. Seetharaman, and K. Palaniappan, "3D patch-based multi-view stereo for high-resolution imagery," in Geospatial Informatics, Motion Imagery, and Network Analytics VIII, vol. 10645. International Society for Optics and Photonics, 2018

Conclusion

- Proposed a novel end-to-end framework for **generation of large scale, realistic, synthetic aerial image sequences** that can be used for evaluation of both 2D and 3D computer vision tasks
- Proposed framework consists of three main modules:
  1. **3D voxel renderer** for view generation,
  2. **deep neural network for image artifact removal**, and
  3. **3D point cloud evaluation** module
- Introduced a novel network **ARNET-Edge** which take in input image along with edge map to learn the structure of the building
- Evaluation of the generated synthetic image sequences for structural similarity to the real images and for utility to produce dense 3D point clouds have shown promising results
- **Comparative evaluations** have demonstrated that the **novel edge-augmented input** and **explicit edge-map processing stream** in the proposed artifact removal network greatly contributes to preservation and recovery of the scene structures