Predicting ground-level scene Layout from Aerial imagery

Muhammad Hasan Maqbool

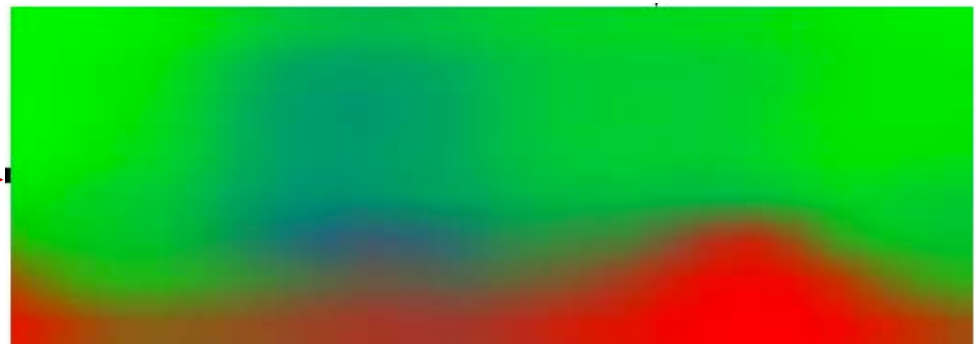
Objective

- Given the overhead image → predict its ground level semantic segmentation
-



Overhead/Aerial Image

Predicted ground level labeling





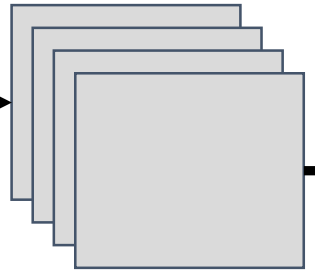
Ground Level Image

Overall Architecture (Training)

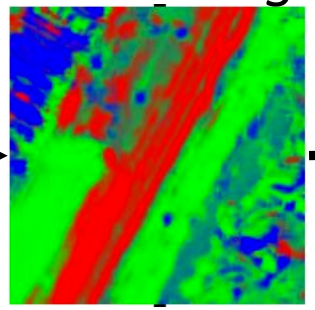


Overhead/Aerial Image

VGG

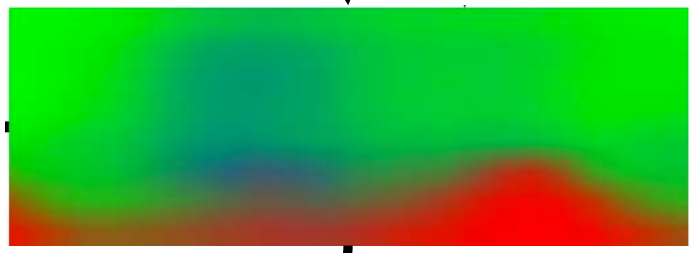


Labeling



Transform

Ground Level Labeling (Predicted)

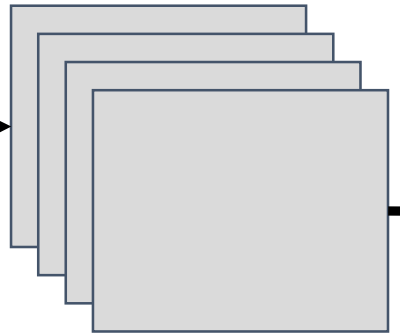


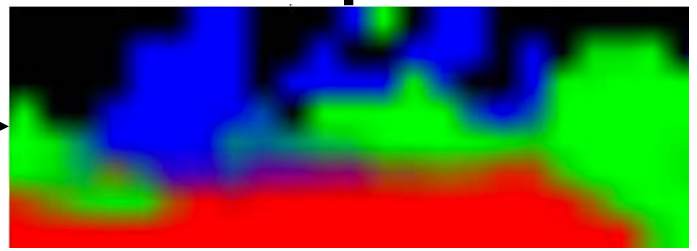
Cross Entropy Loss



Ground Level Image

SegNet





Ground Level Labeling

Ground Truth of Ground-level Segmentation

1. Generate the segmentation of ground level image using (pre-trained model)
    1. SegNet → This gives “Ground Truth” for training
2. To avoid manual pixel annotation of their new dataset (overhead imagery)

Algorithm Overview

Aerial Image to Ground Level (Predicted) Segmentation

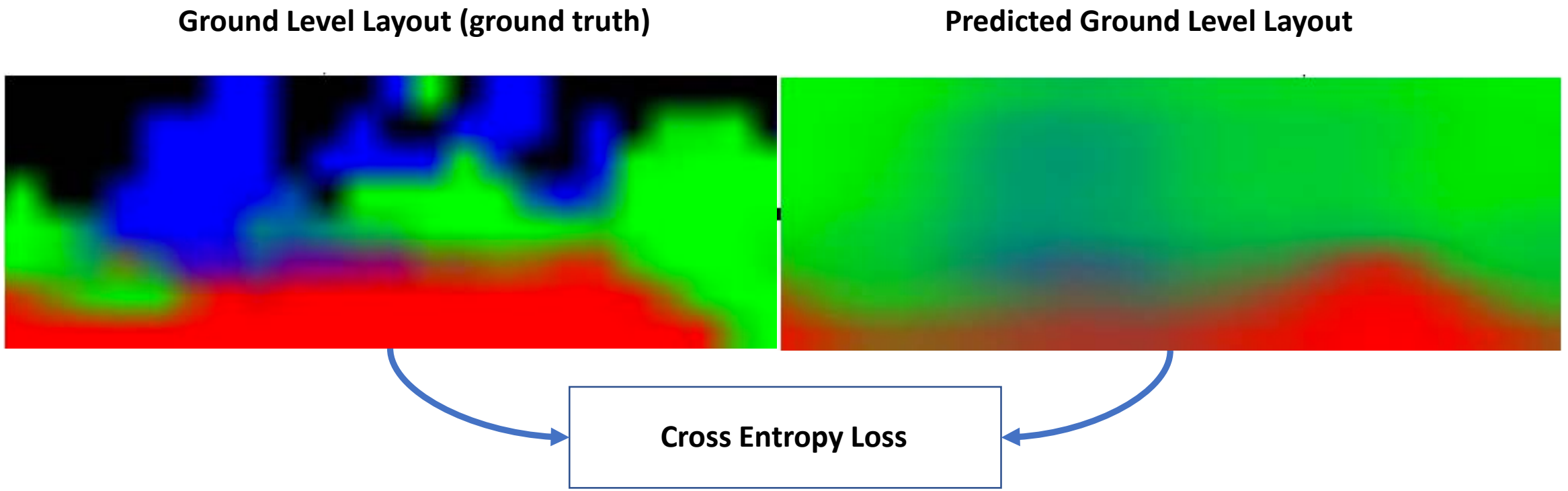


128 Conv 1 x 1
64 Conv 1 x 1
1 Conv 1 x 1

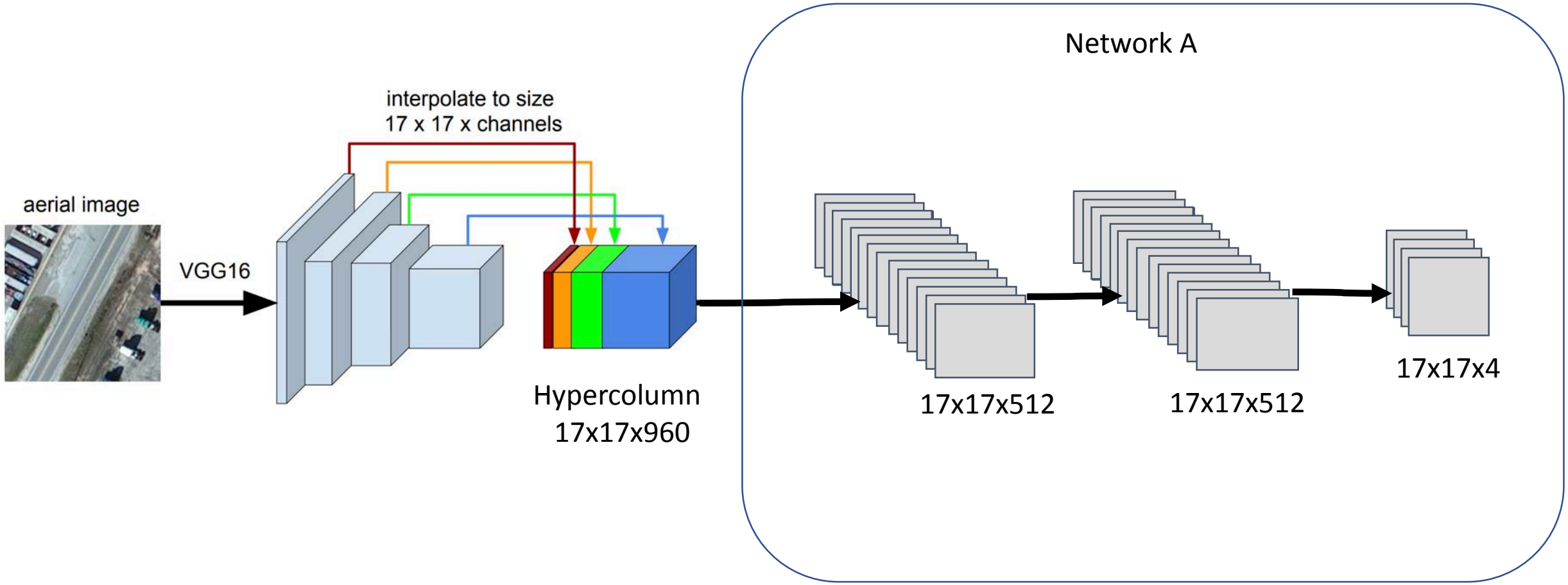
Algorithm Overview

Predicted and Ground Truth Layouts Distance Minimization

- Minimize the pixel distance between the ground layout and the aerial-to-ground layout prediction



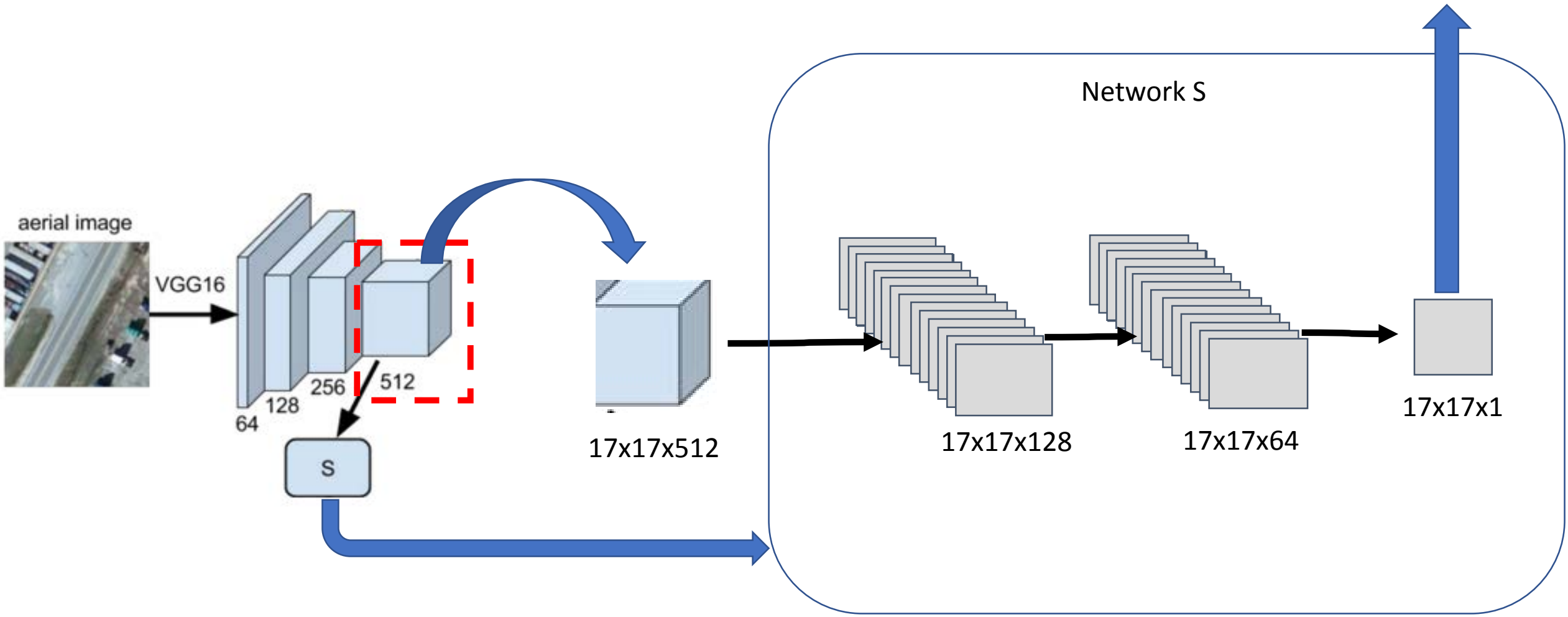
Network A



Three conv layers employ 1x1 convolution
Reduce channel dimension from 960 to 4

Network S

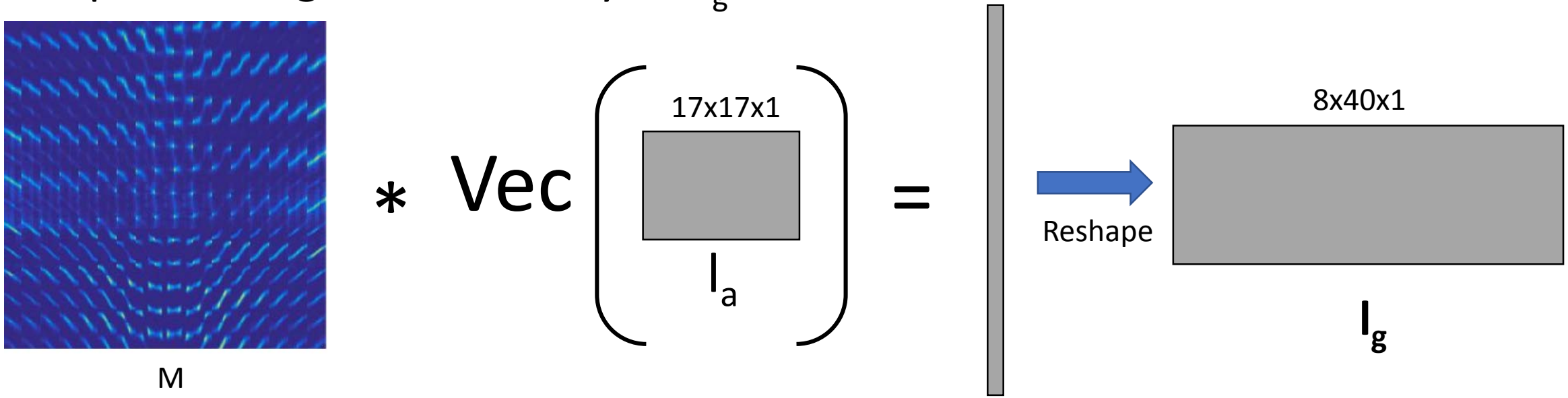
Encode the global information of the aerial image



Three conv layers employ 1x1 convolution
Reduce channel dimension from 512 to 1

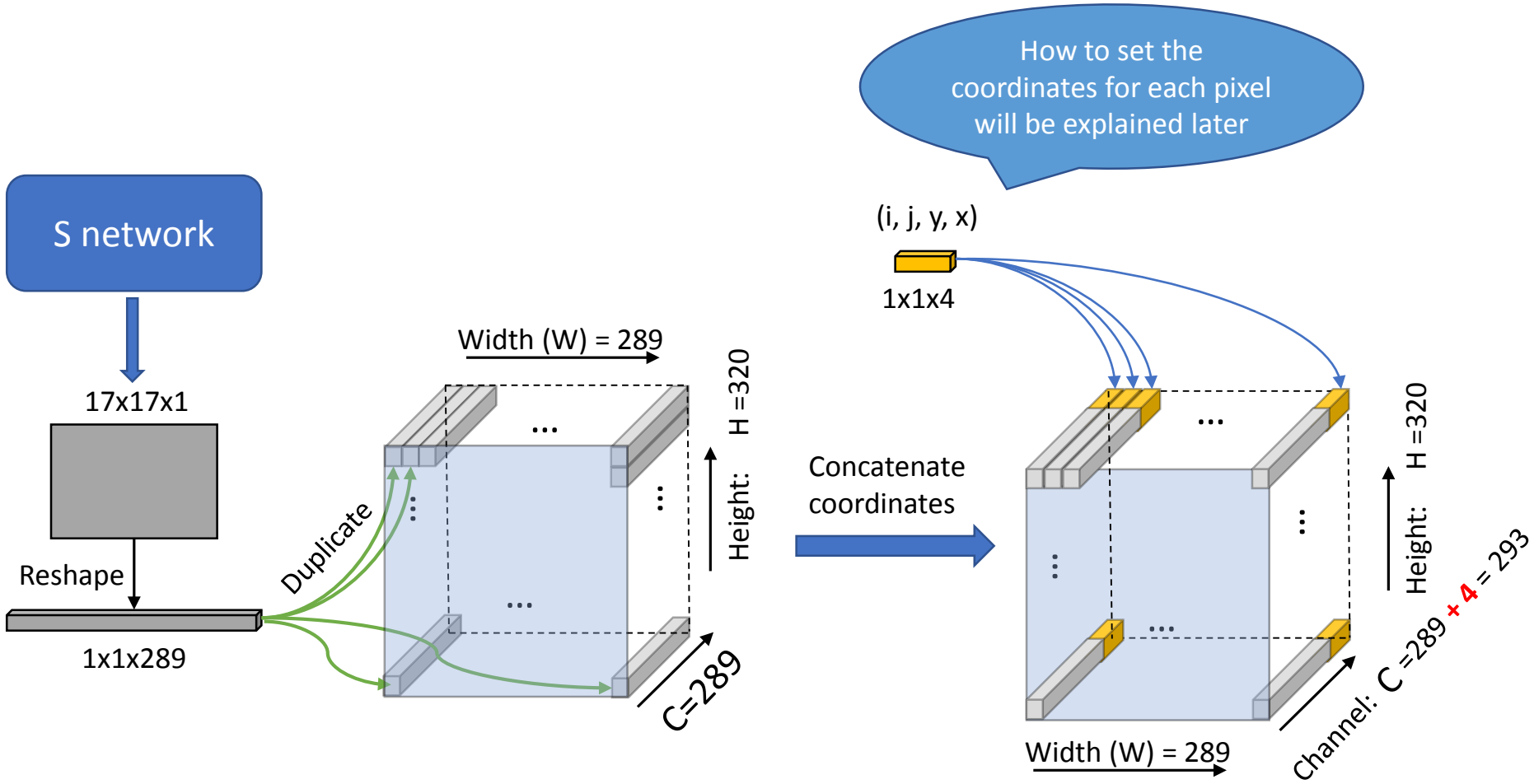
Determine the Size of the Transformation Matrix

- Aerial image layout l_a is multiplied by transformation matrix *M* to get the predicted ground-level layout l_g

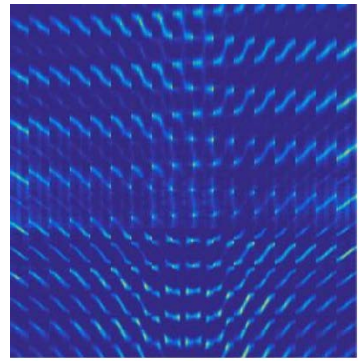


- Based on the input size and the desired output size, we can determine the size of *M -----> In this case, we have M = 320x289*

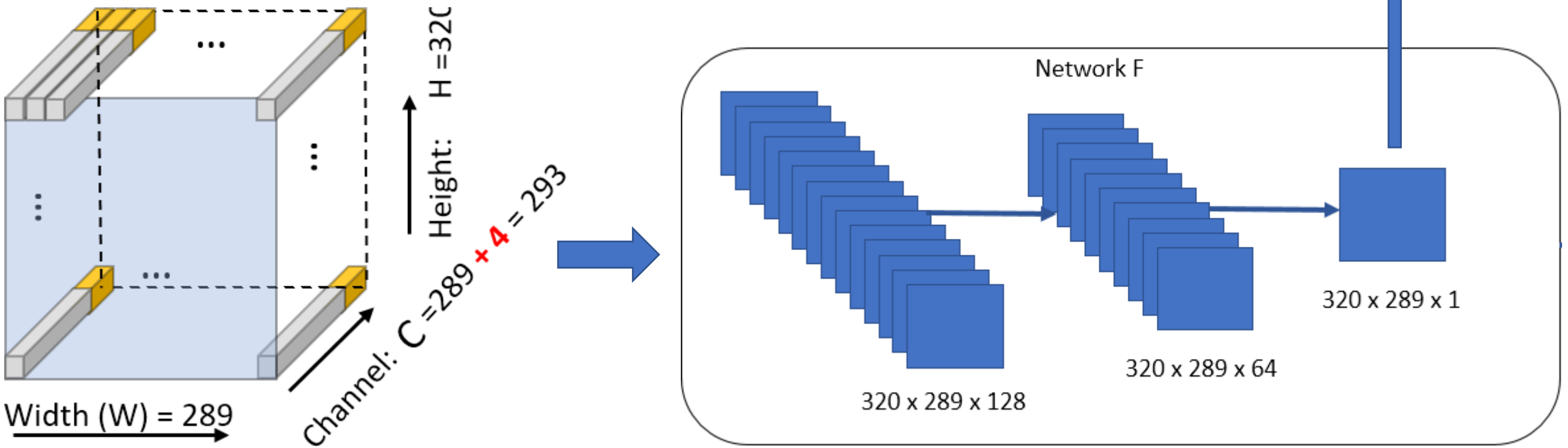
Construct Input to Network F



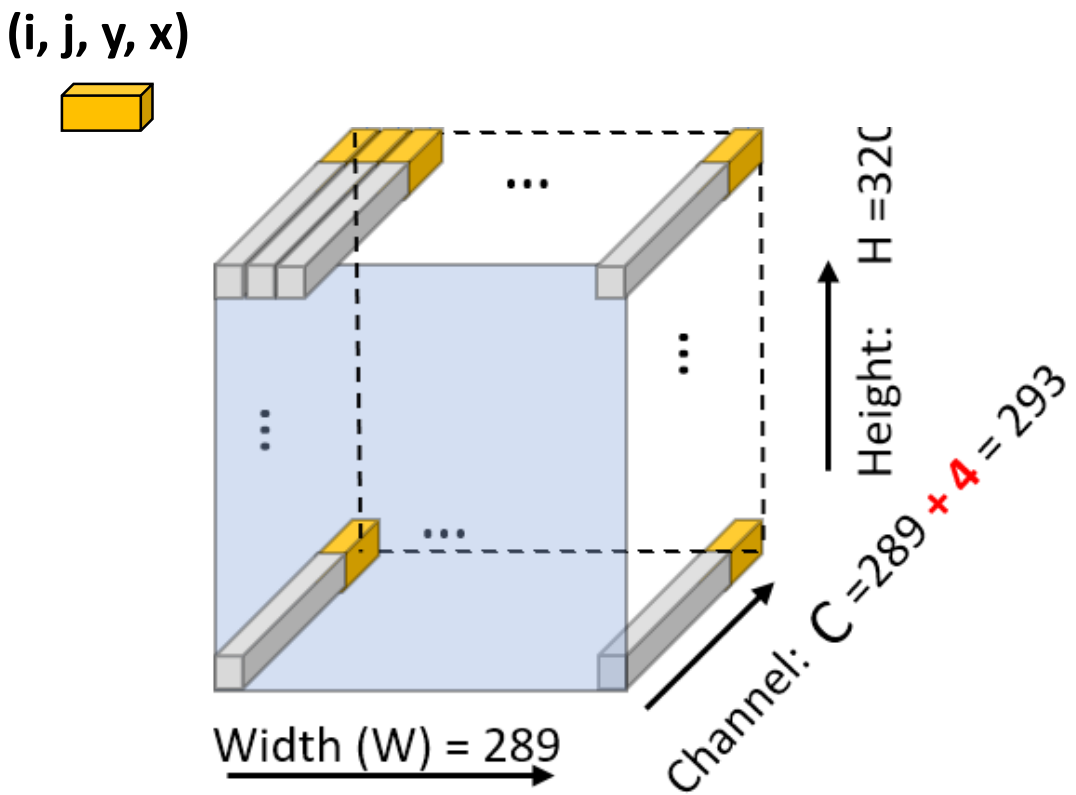
Network F



Transformation matrix

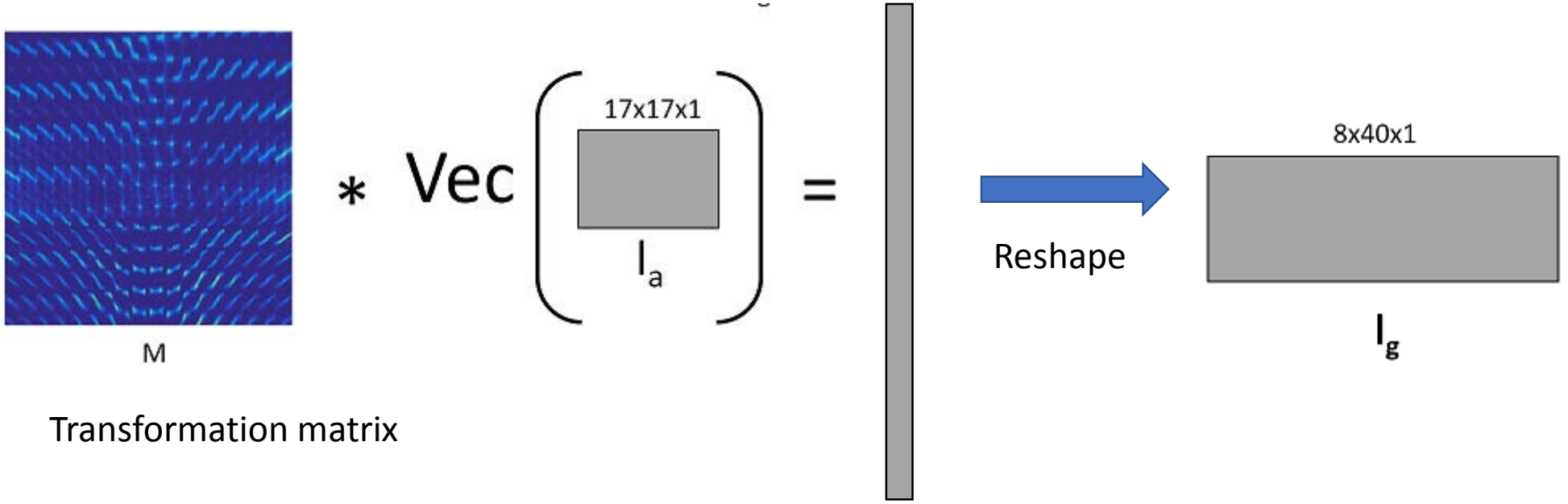


How to determine the coordinates?

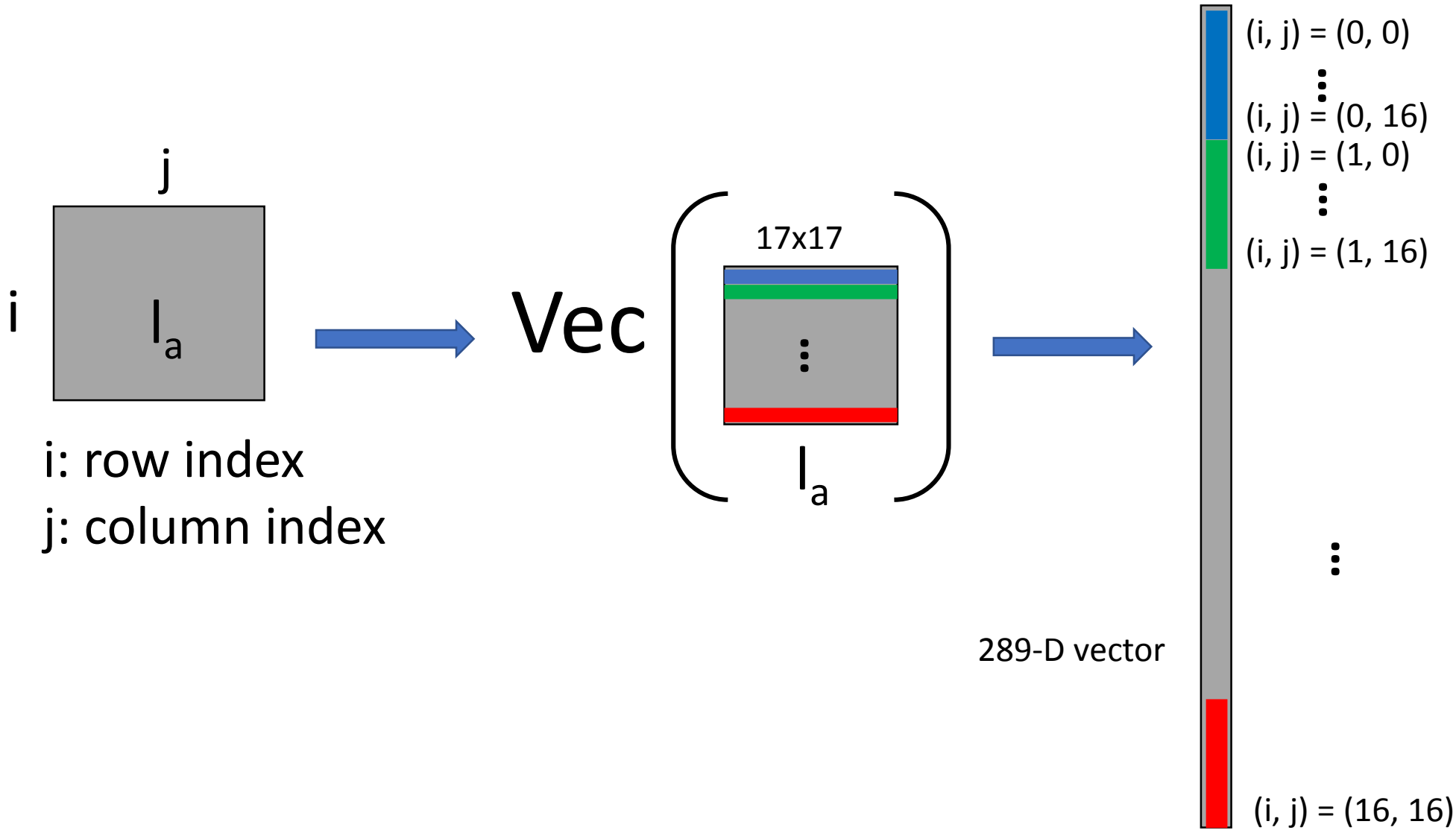


How to determine the coordinates?

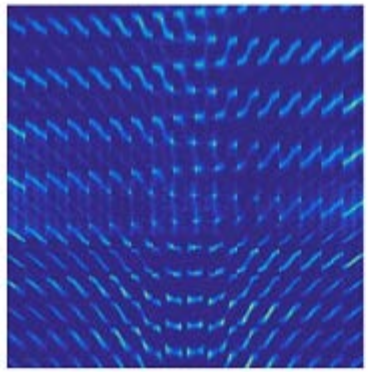
Recall – the following procedure is used to get the predicted ground-level layout



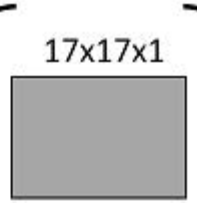
How to determine the coordinates?



How to determine the coordinates?



M

$*$ Vec $\left(\right.$  $\left.\right) =$

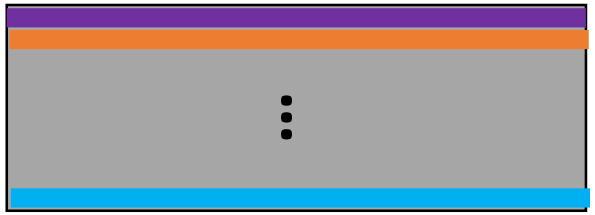
17x17x1

I_a



320-D vector

Reshape



x

8x40

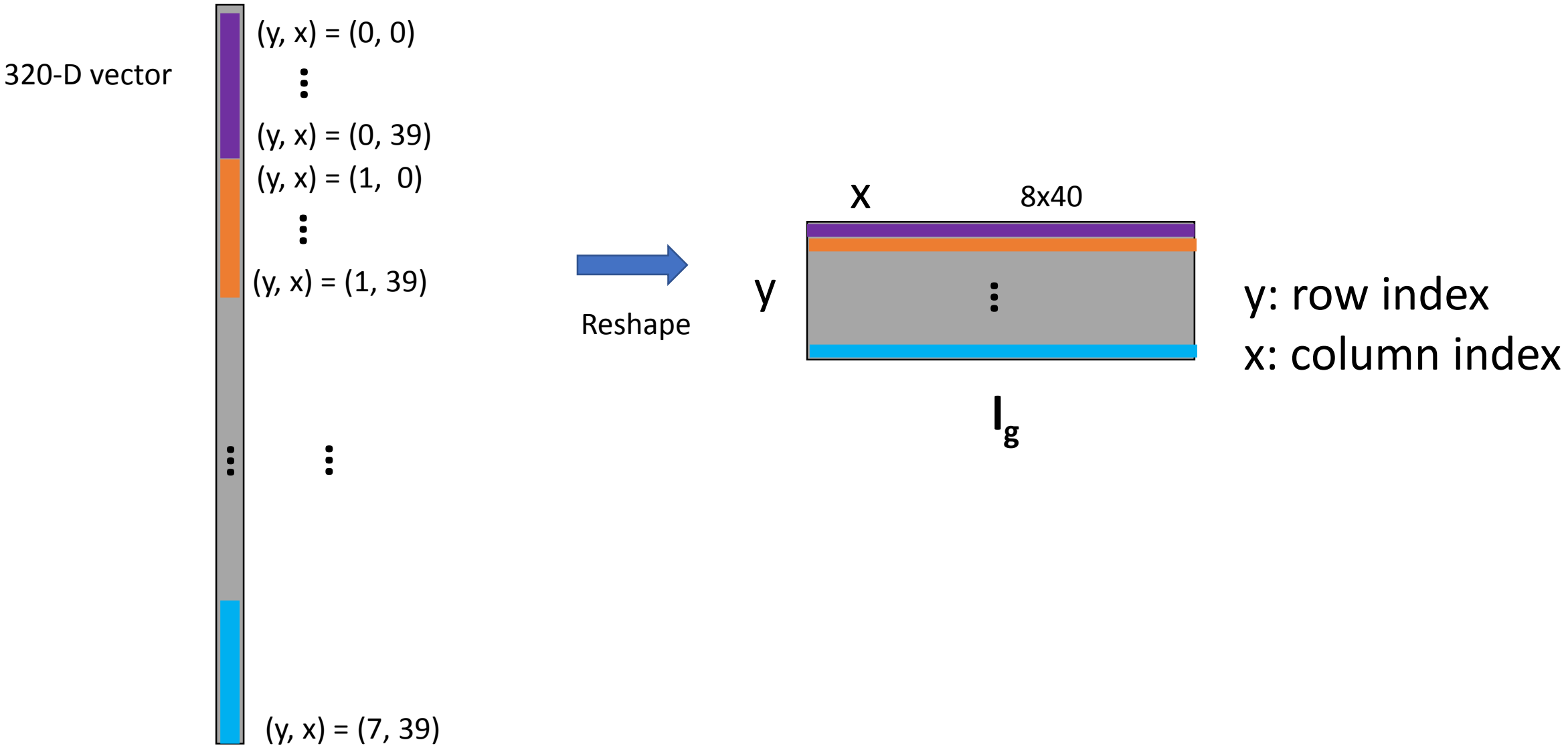
y

I_g

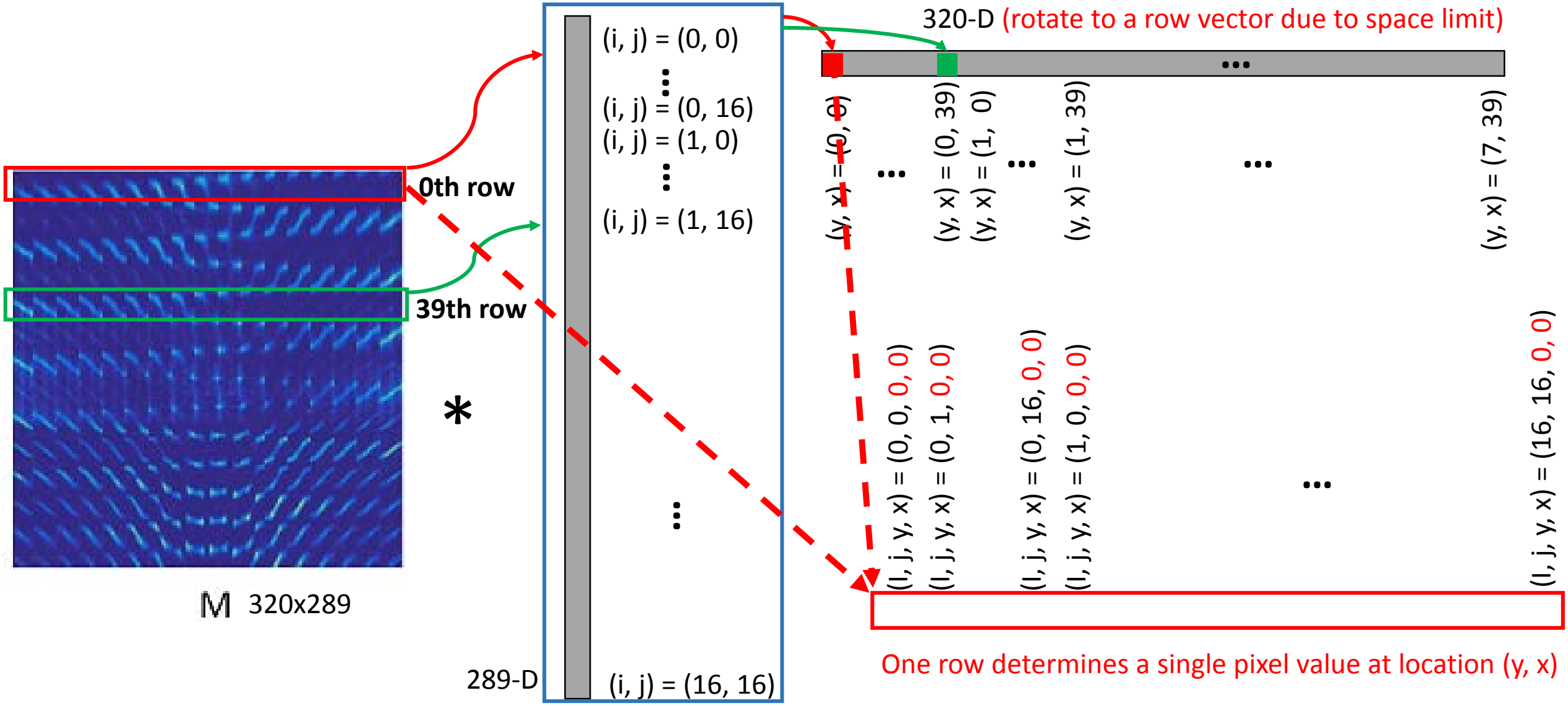
y: row index

x: column index

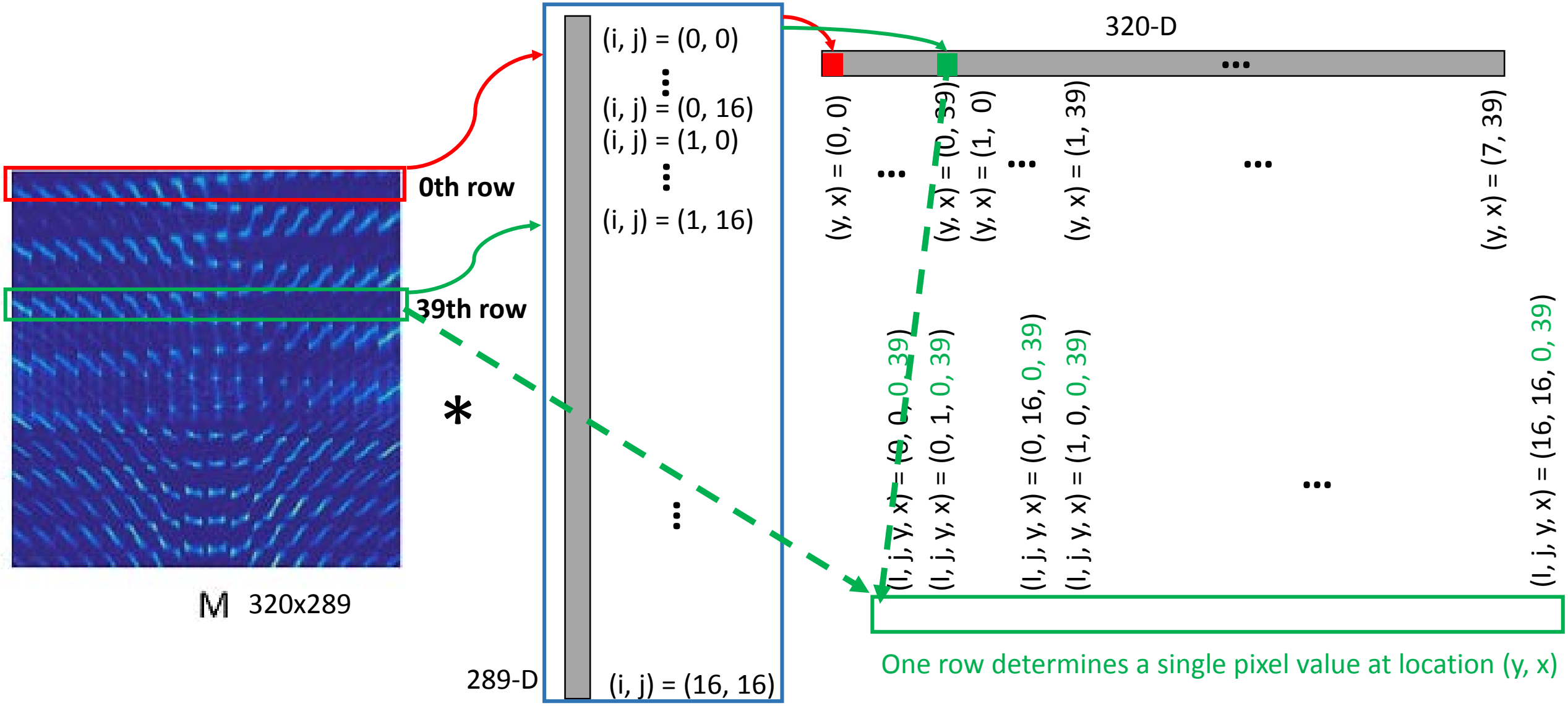
How to determine the coordinates?



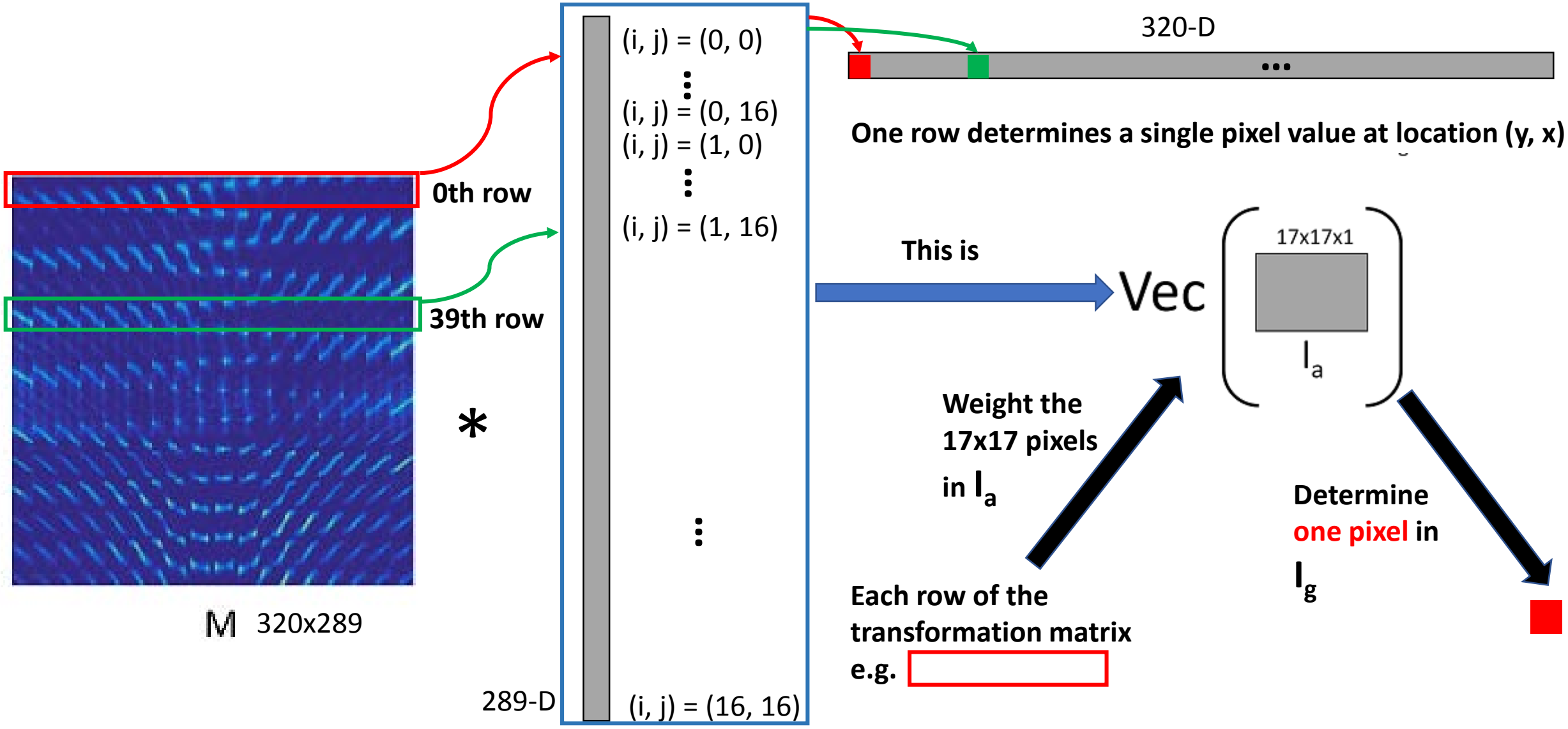
How to determine the coordinates?



How to determine the coordinates?

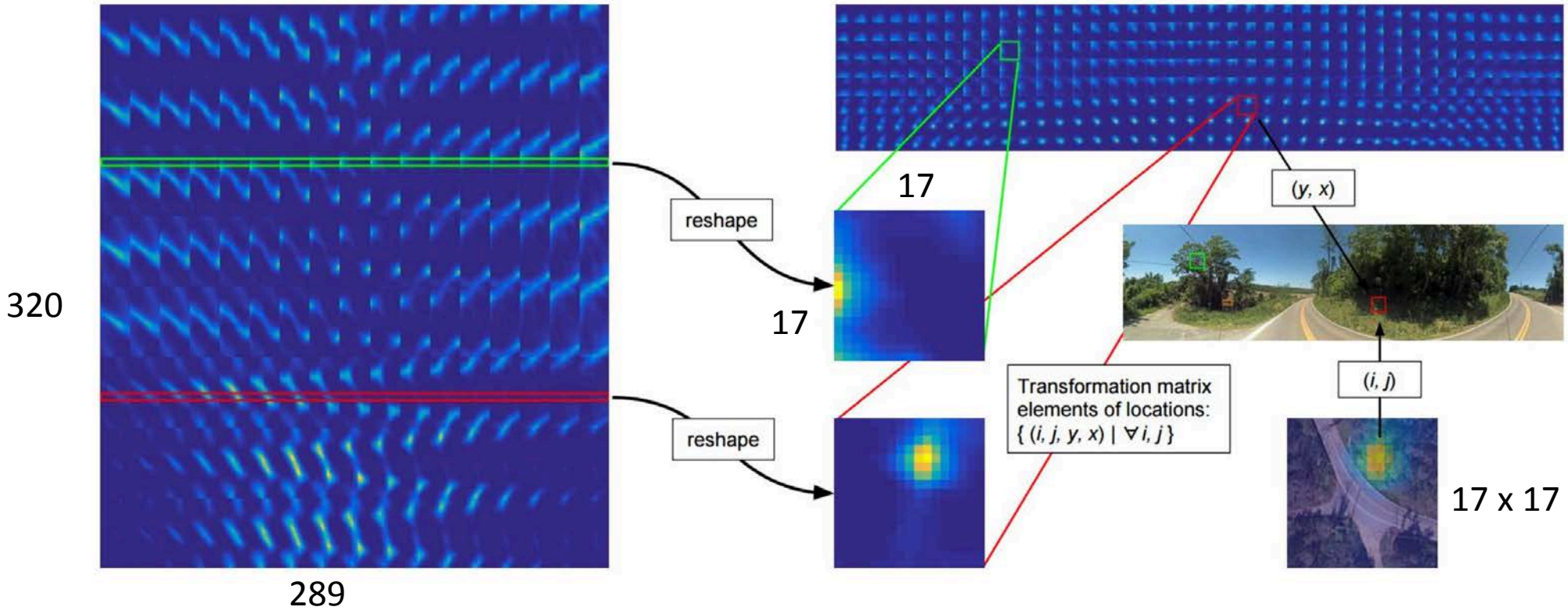


Interpret the transformation matrix



Transformation Matrix Visualization

- The transformation matrix encodes the relationship between the aerial-level pixels and the ground-level pixels



Dataset

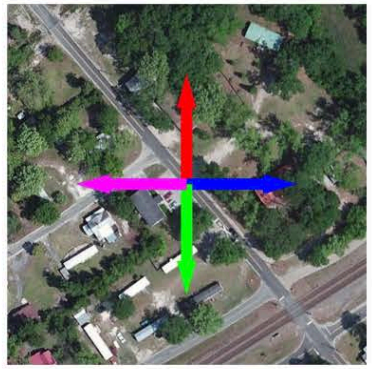
Cross-View Dataset

- CVUSA contains
- approximately 1.5 million geo-tagged pairs of ground and aerial images from across the United States.
- Google Street View panoramas of CVUSA as our ground level images
- Corresponding aerial images at zoom level 19
- 35,532 image pairs for training and 8,884 image pairs for testing.

Cross-View Dataset: An Example

In the aerial images, north is the up direction.

In the ground images, north is the central column.

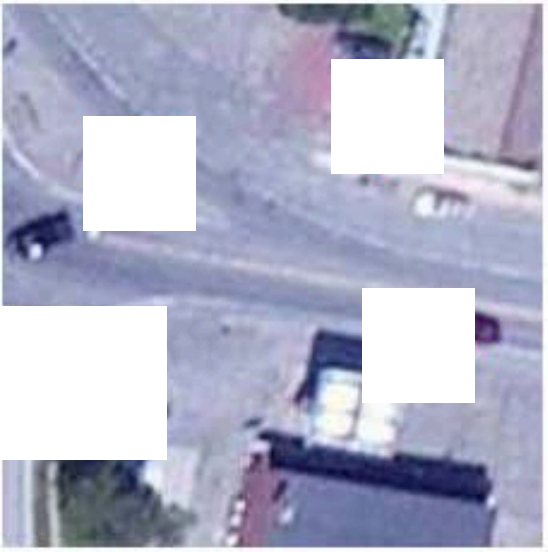


Aerial



Ground

Cross-View Dataset





Cross-View Dataset

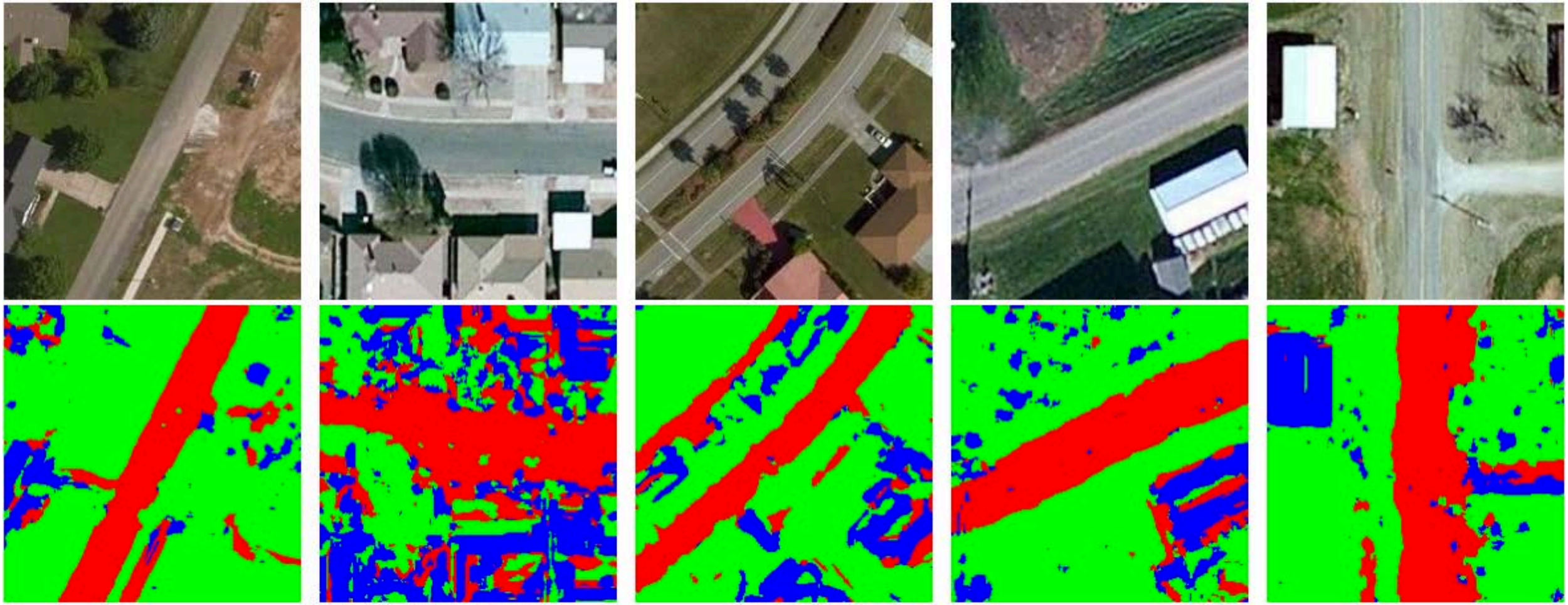




Evaluation and Applications

Weakly Supervised Learning

Weakly Supervised Learning for Aerial Image Classification and Segmentation

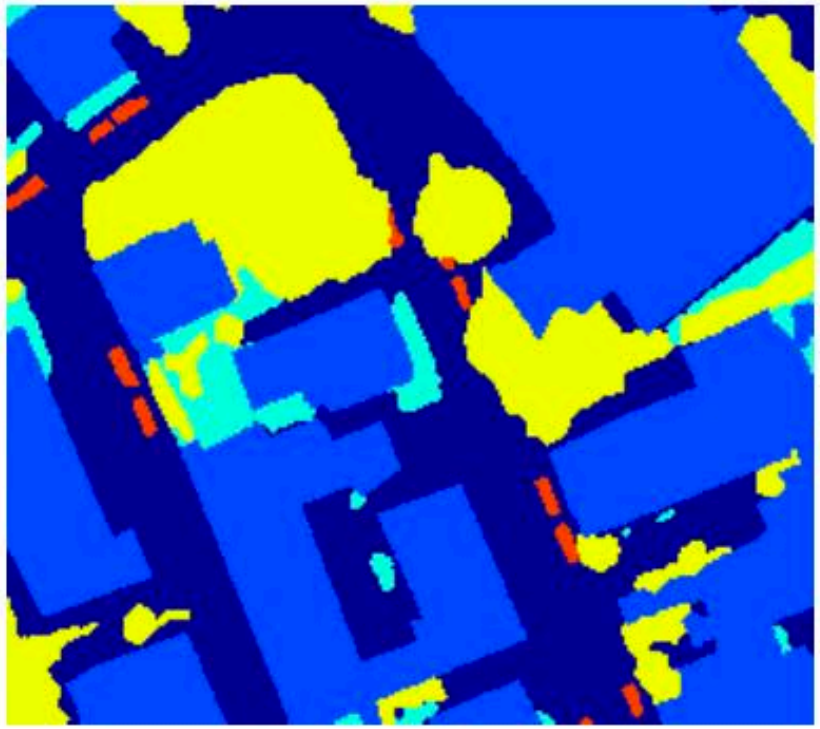


As Pre-trained Model

Network Pre-training for Aerial Image Segmentation





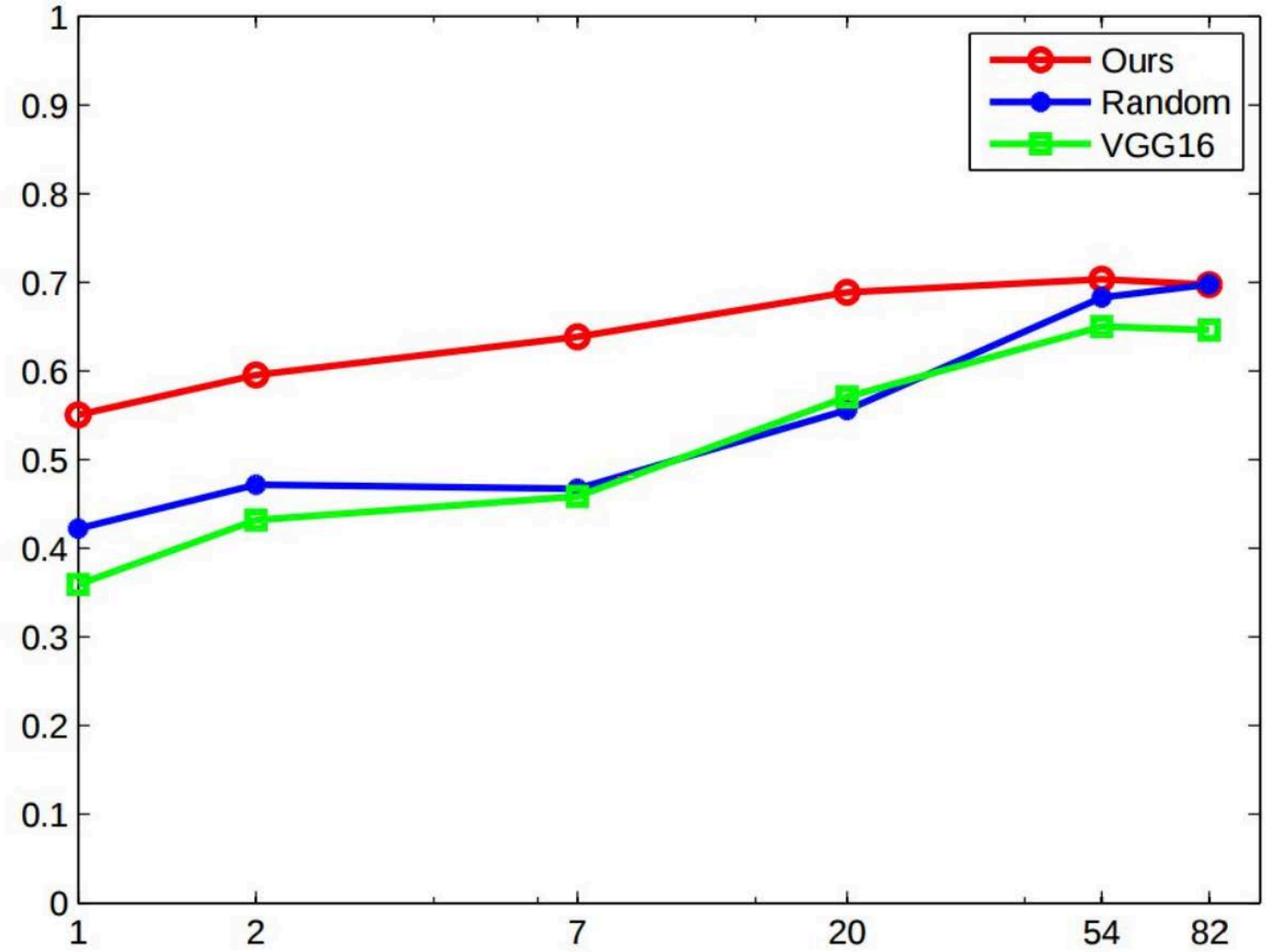


Network Pre-training for Aerial Image Segmentation

Per-class Precision on the ISPRS Segmentation Task

Class	Init.	Number of training samples					
		1	2	7	20	54	82
Imp.	Ours	0.67	**0.74**	**0.63**	**0.64**	0.66	0.64
	Random	**0.70**	0.70	0.54	0.62	0.61	**0.73**
	VGG16	0.60	0.55	0.55	0.61	**0.70**	0.59
Bldg	Ours	0.72	**0.76**	**0.76**	**0.80**	0.75	**0.78**
	Random	0.56	0.62	0.63	0.64	**0.82**	0.71
	VGG16	**0.78**	0.72	0.71	0.69	0.70	0.75
Low.	Ours	**0.37**	**0.43**	**0.51**	**0.65**	**0.67**	**0.67**
	Random	0.29	0.29	0.29	0.37	0.67	0.64
	VGG16	0.25	0.25	0.29	0.44	0.53	0.57
Tree	Ours	**0.68**	**0.54**	**0.71**	**0.71**	**0.74**	0.74
	Random	0.42	0.46	0.49	0.56	0.71	0.69
	VGG16	0.36	0.44	0.50	0.55	0.65	**0.74**
Car	Ours	**0.13**	**0.46**	**0.67**	**0.48**	**0.48**	0.49
	Random	0.05	0.08	0.10	0.25	0.45	**0.57**
	VGG16	0.05	0.11	0.20	0.20	0.25	0.23

Network Pre-training for Aerial Image Segmentation



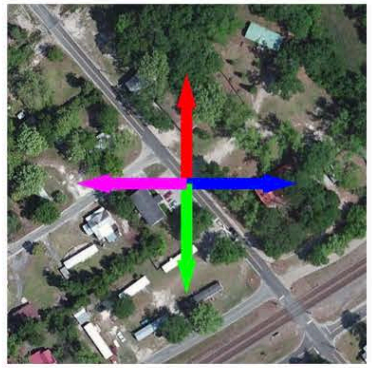
Orientation Estimation

Orientation Estimation

- Given a street view image $I_{g,}$ estimate its orientation
  1. Get the corresponding aerial image (using GPS tag), I_a
  2. Infer the semantic labeling of the query image, $I_g \rightarrow L_g$
  3. Predict the ground image labels from the aerial image using the proposed network, I_a -> Lg'
  4. Compute the cross entropy between Lg and Lg' in a sliding window fashion across all possible orientations
  5. The smallest entropy is the optimal estimated orientation

Orientation Estimation

- Their CVUSA dataset contains aligned image pairs





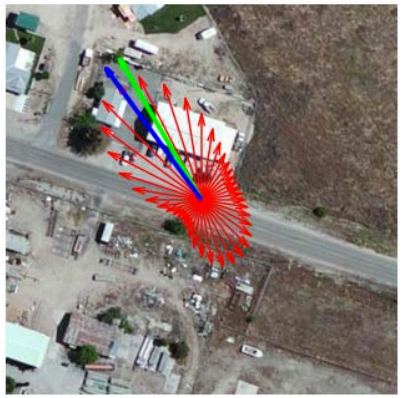
- Therefore, the learned network predicts the ground (panorama image) layout in a fixed orientation – served as reference

Orientation Estimation

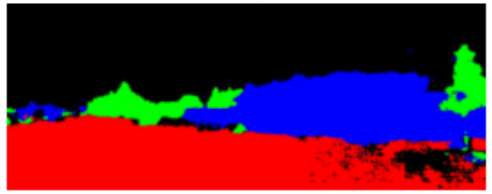
Query street view image



Get aerial image



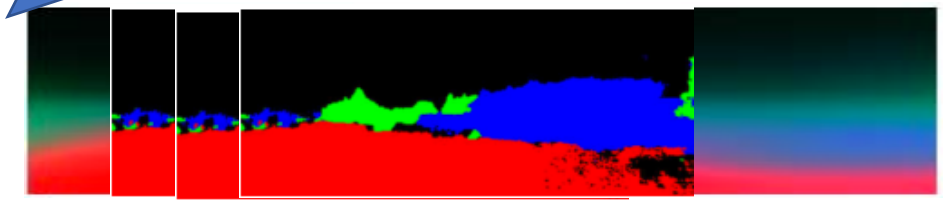
Get segmentation
(e.g. SegNet)



Get predicted ground-level layout using the proposed network

Compute cross-entropy in sliding window fasion

W

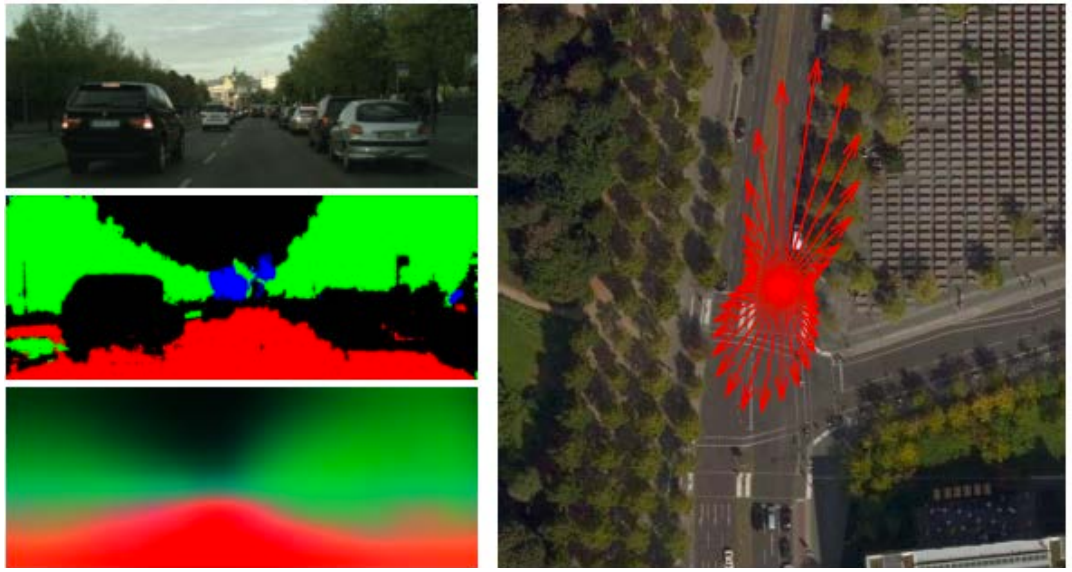


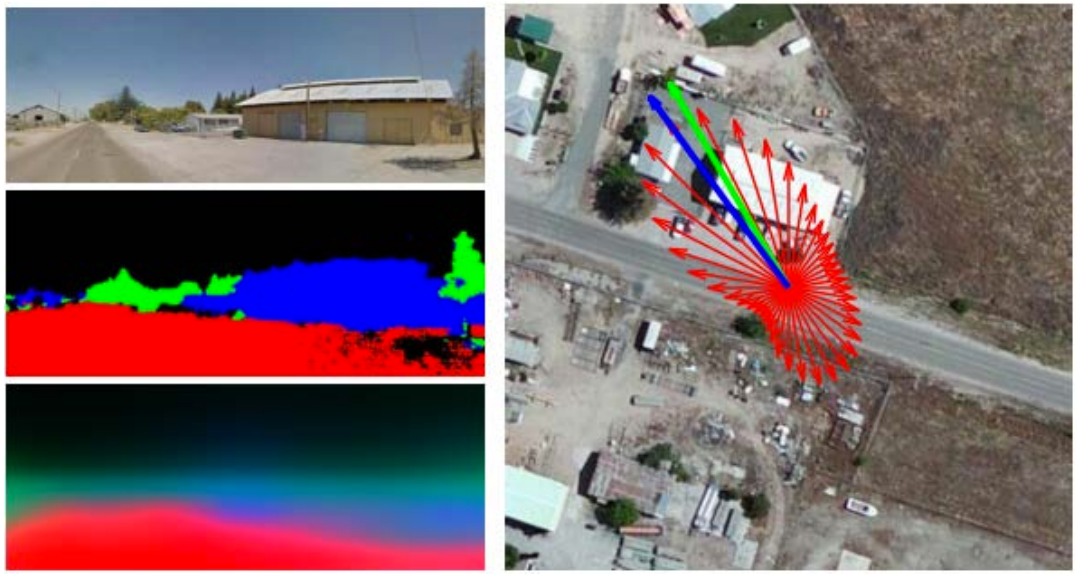
H

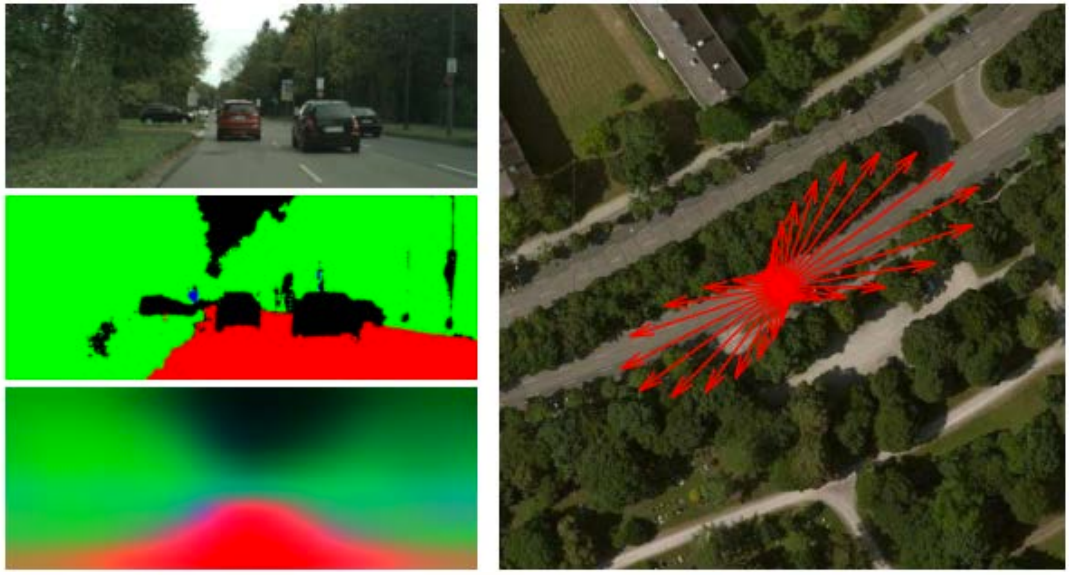
1 column = 360 degree / W

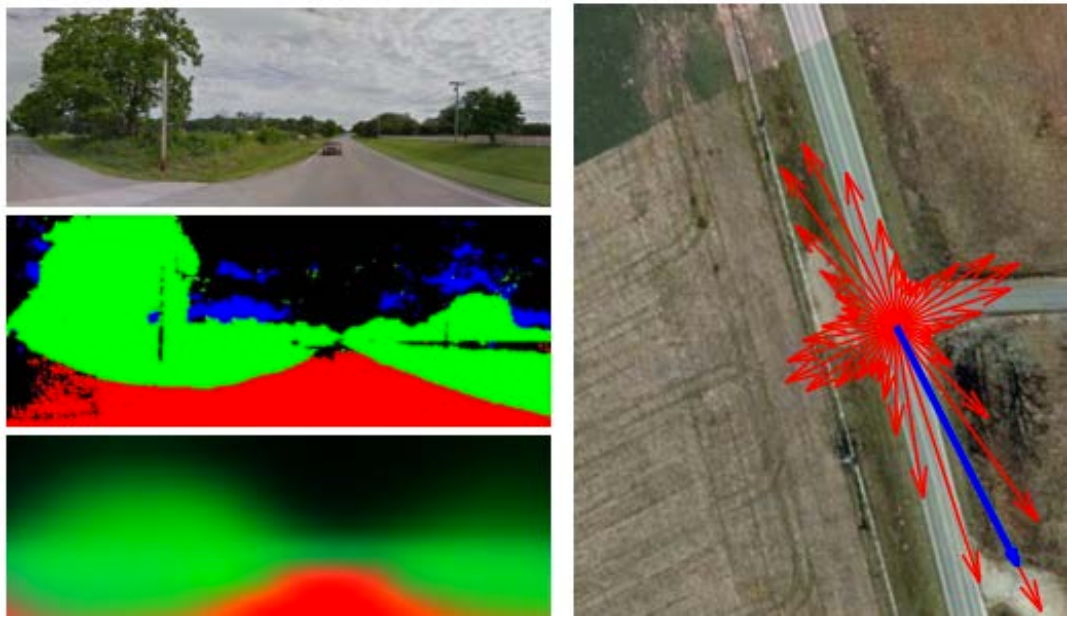
Predicted ground-level layout of the panorama image

Orientation Estimation



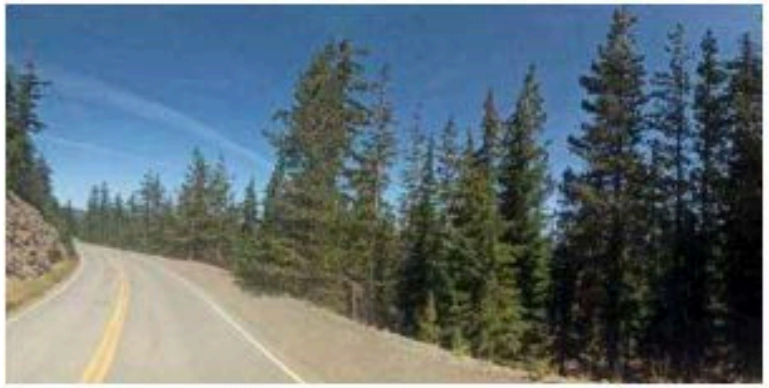


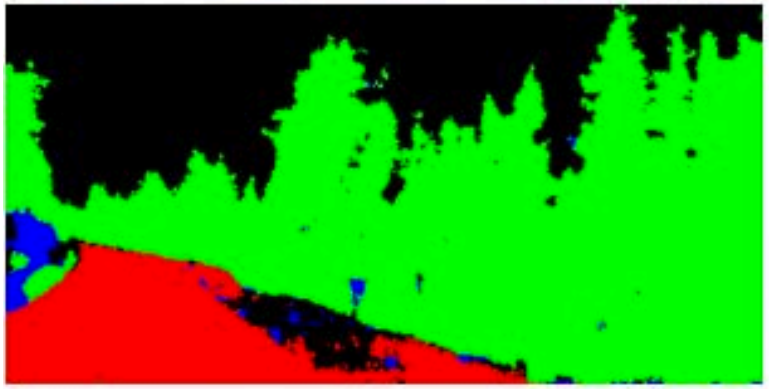




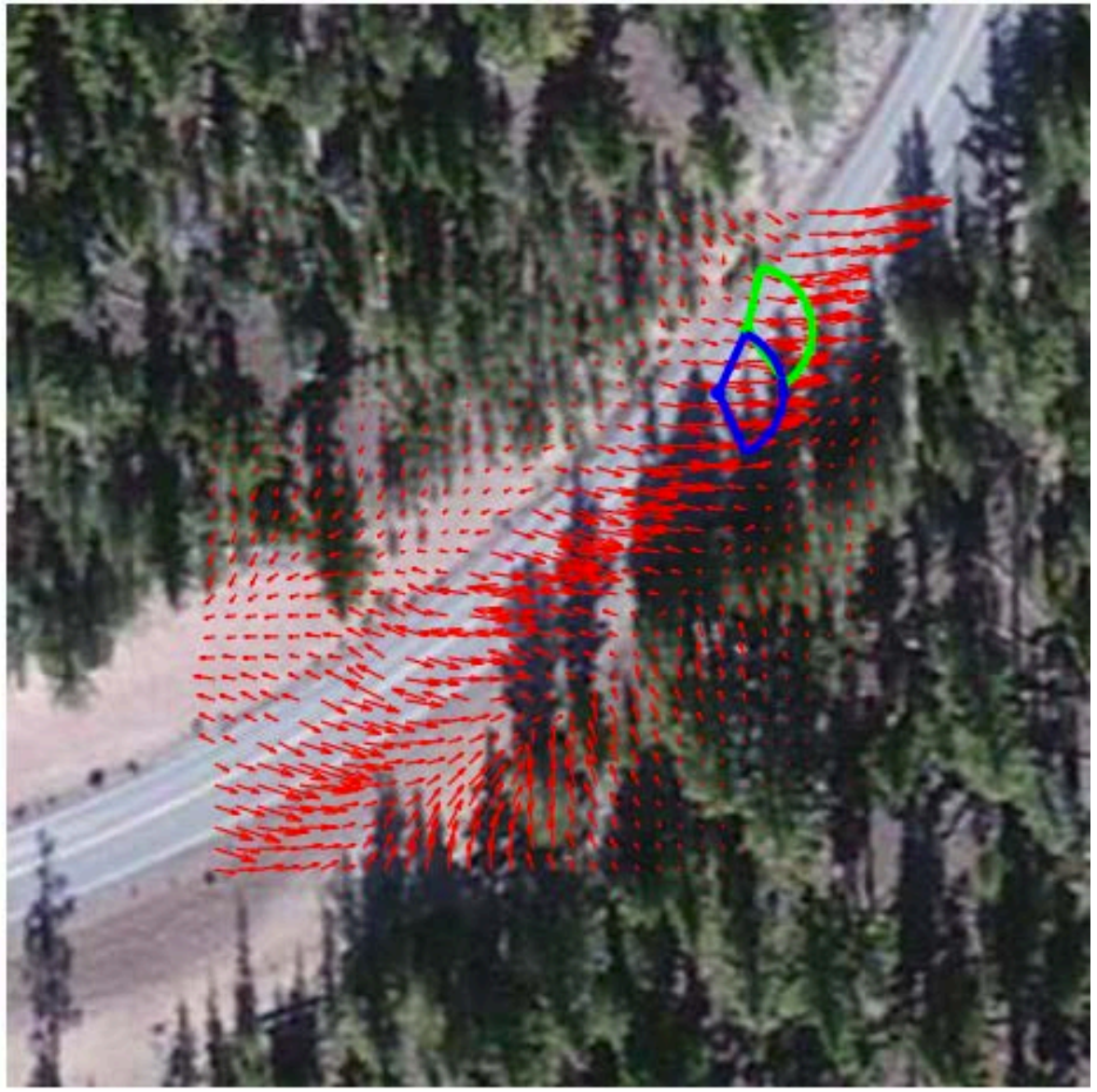
Geocalibration

Fine-Grained Geocalibration



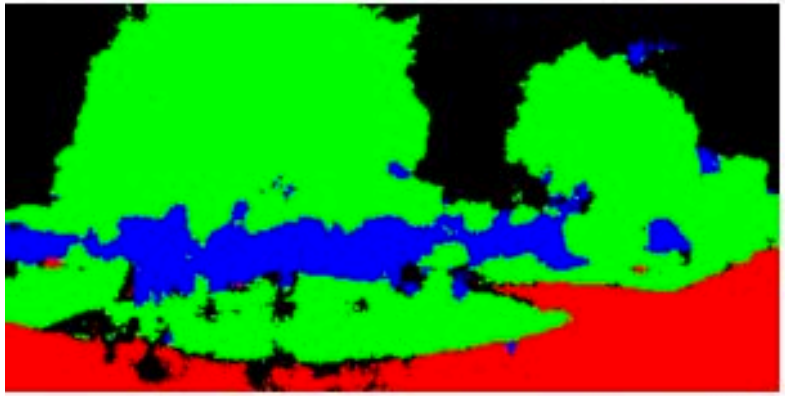


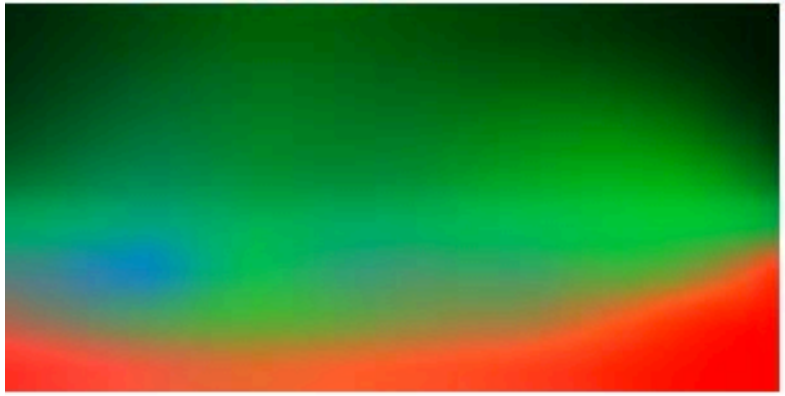


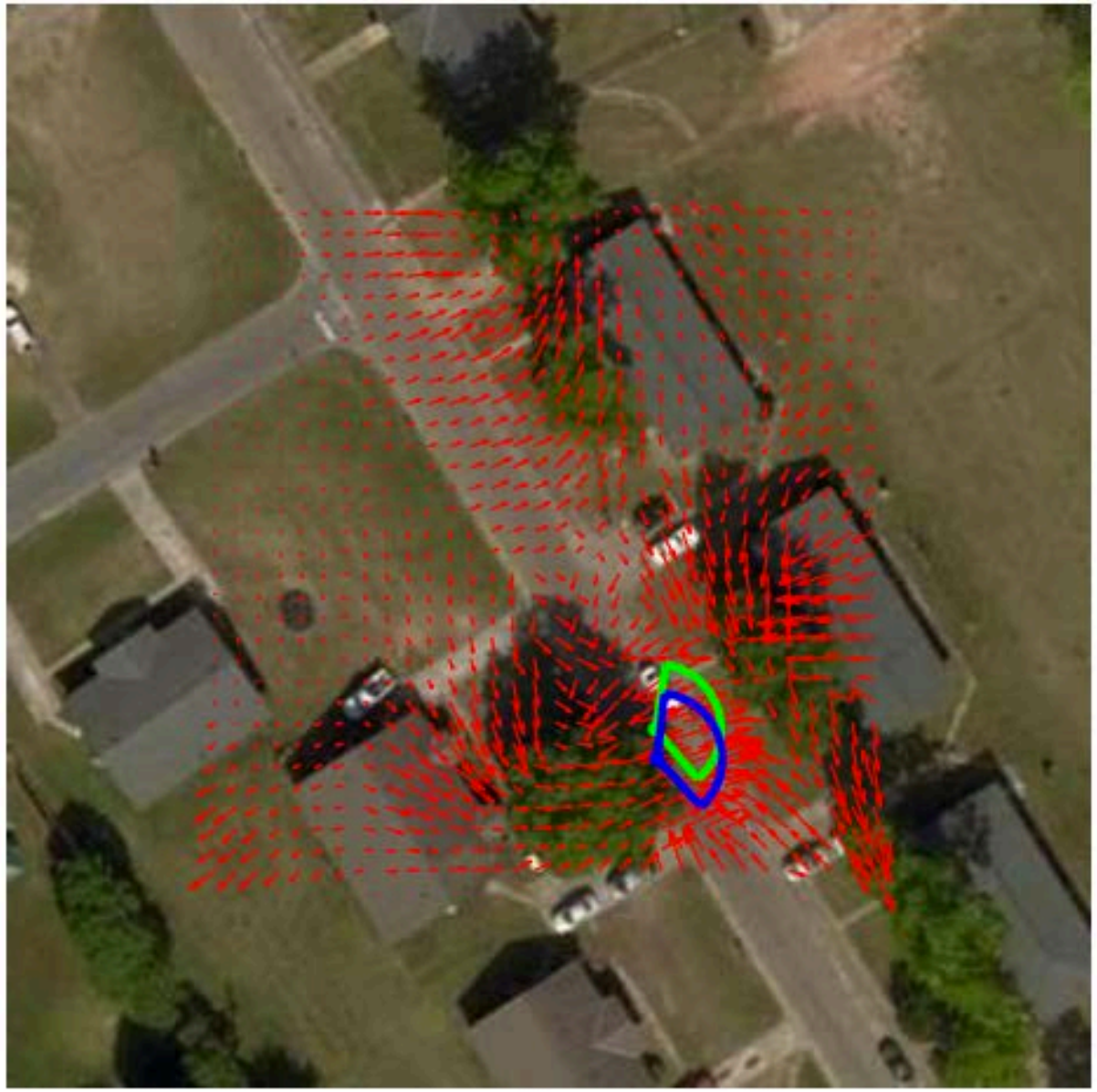


Fine-Grained Geocalibration









Ground Level Image Synthesization

Ground-level Image Synthesization







Ground-level Image Synthesization







Conclusion

- A novel strategy to relate aerial-level imagery and the ground-level imagery
- A novel strategy to exploit the automatically labeled ground images as a form of weak supervision for aerial imagery understanding
- Show the potential of this method in geocalibration and ground appearance synthetization

Questions