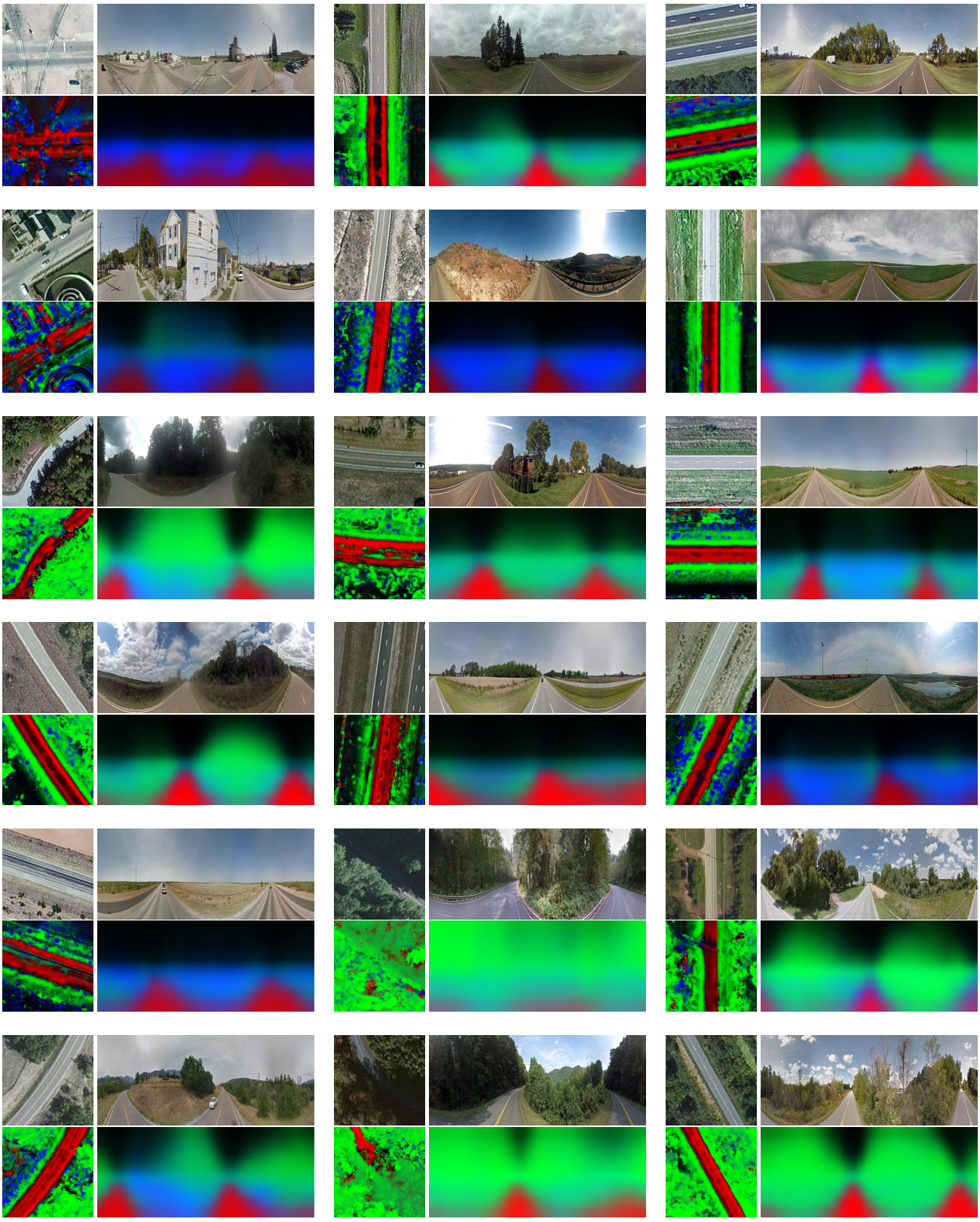
Supplementary Material : Predicting Ground-Level Scene Layout from Aerial Imagery

Menghua Zhai ted@cs.uky.edu
Zachary Bessinger zach@cs.uky.edu
Scott Workman scott@cs.uky.edu
Nathan Jacobs jacobs@cs.uky.edu
Computer Science, University of Kentucky

Overview

This supplementary material contains additional details and qualitative results from our experiments. Figure 1 shows randomly selected qualitative results from our weakly supervised learning task. Figure 2 shows additional fine-grained geocalibration results. Table 1 and Table 2 describe the complete network structure for the ground image synthesis application and additional qualitative results are shown in Figure 3.



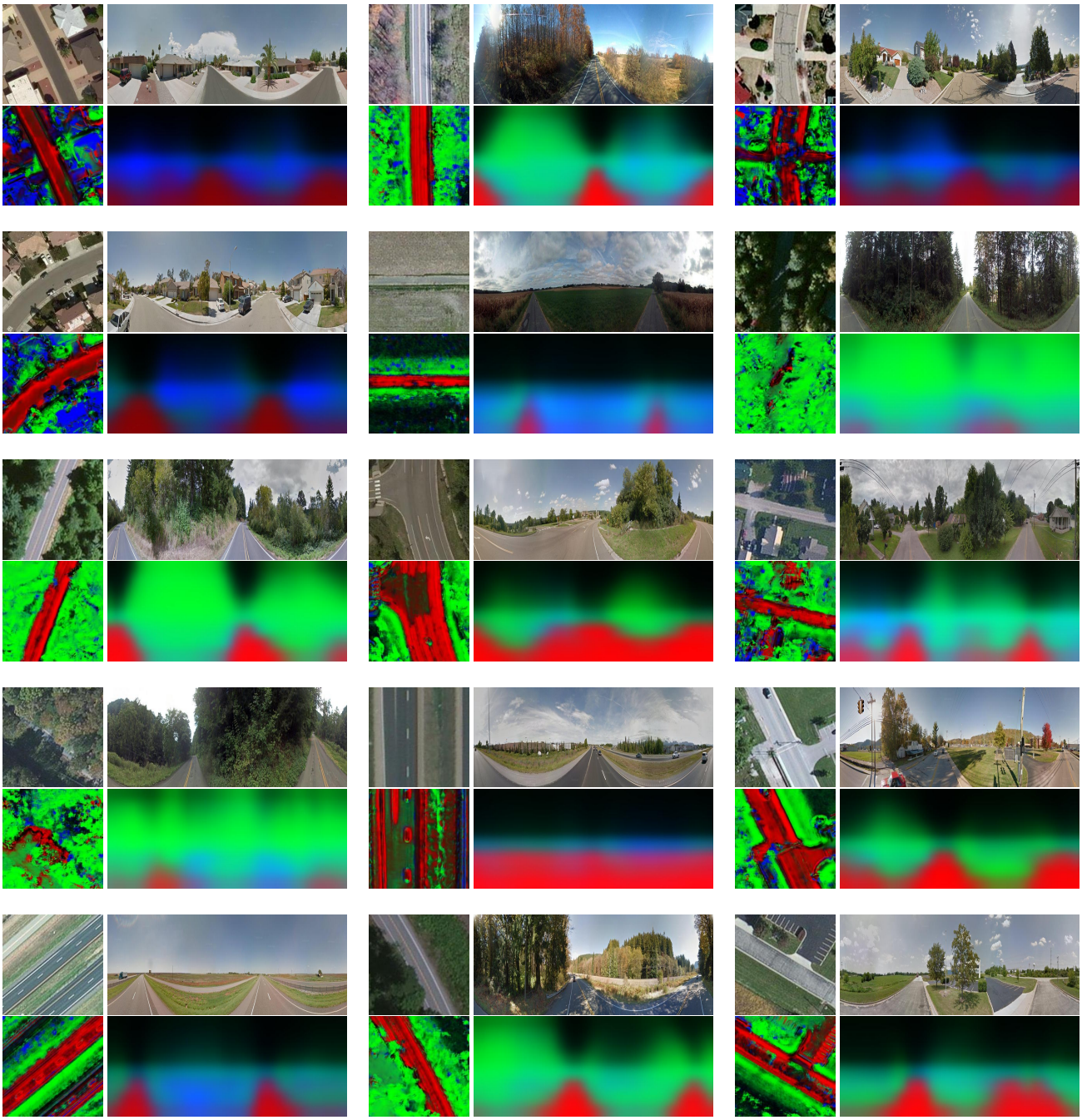
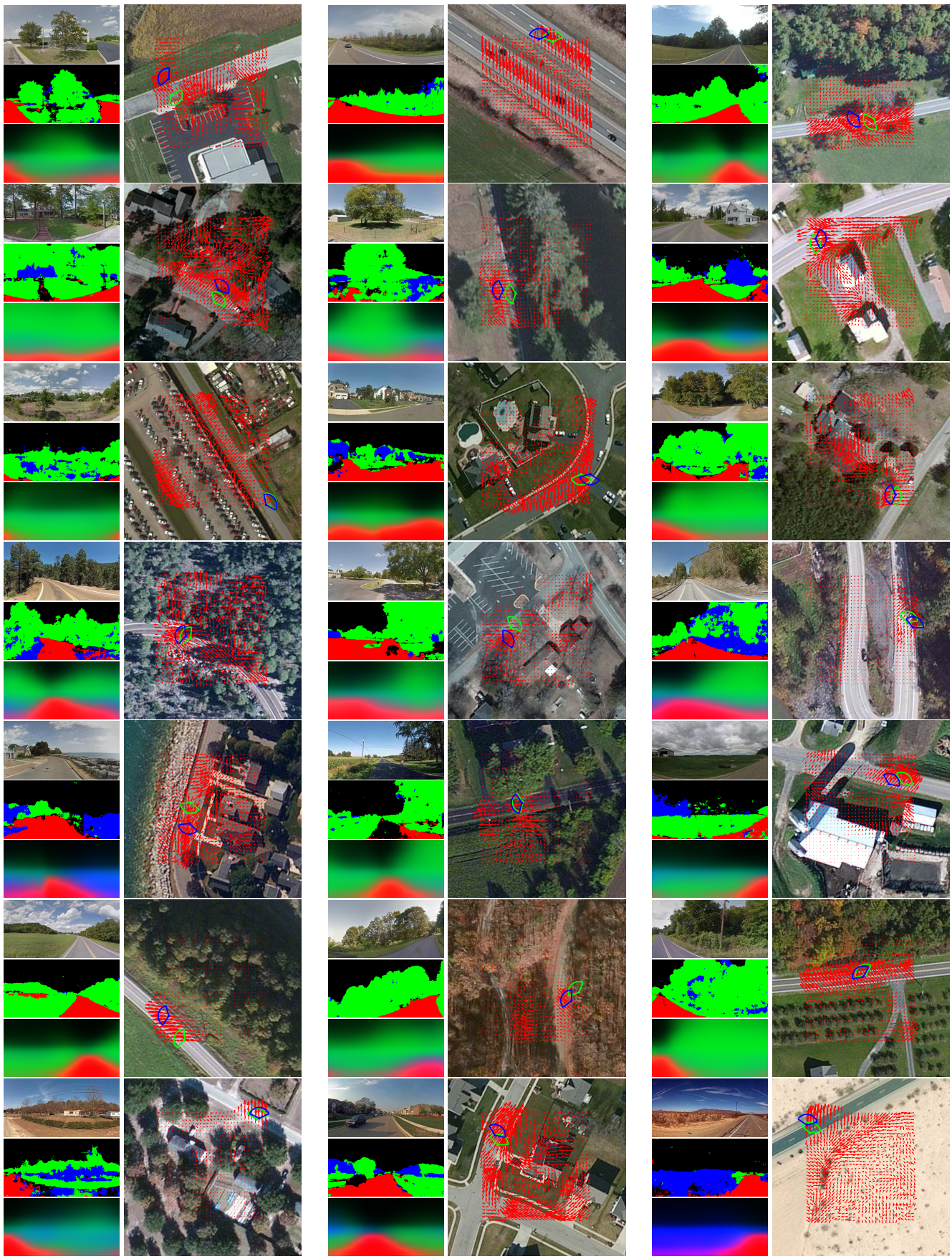


Figure 1. Randomly selected test image results from our weakly supervised learning method. The left column shows the aerial image (top) and the corresponding pixel-level labeling (bottom); The right column shows the ground image (top) of the same location and its pixel-level labeling (bottom) inferred by our model from the aerial image pixel-level labeling. We visualize three classes: *road* (red), *vegetation* (green), and *man-made* (blue).



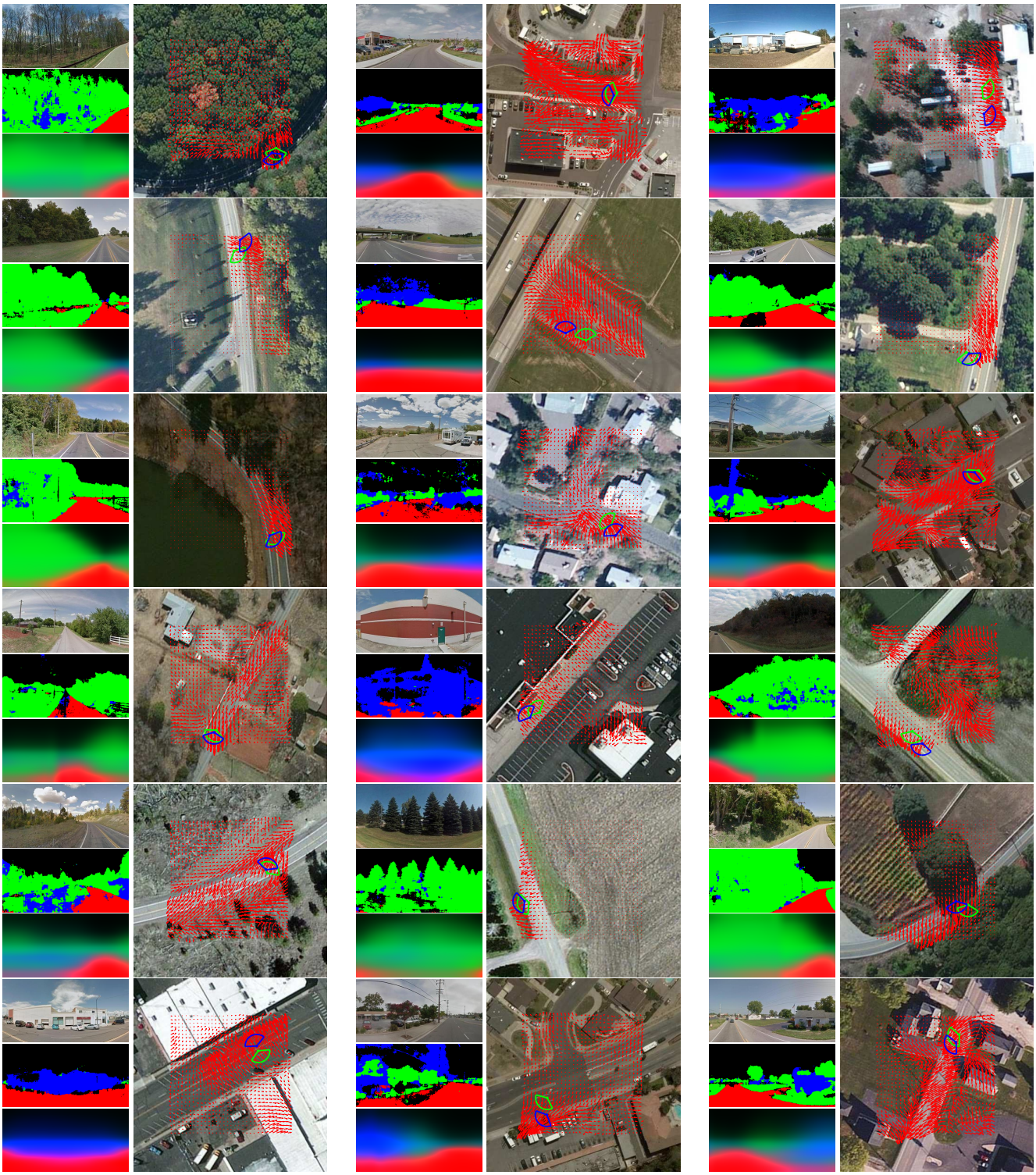


Figure 2. Additional fine-grained geocalibration results. (left) Visualized from top to bottom are I_g, L_g, and $L_{g'}$ respectively. We visualize three classes: *road* (red), *vegetation* (green), and *man-made* (blue). (right) Orientation flow map (red), where the arrow direction indicates the optimal direction at that location and length indicates the magnitude. We also show the optimal prediction and the ground-truth frustums in blue and green respectively.

Table 1. Deep generator network architecture. All deconvolutions use a stride of 2. f is the extracted cross-view feature, and z is Gaussian noise.

Input	Input Shape	Operation	Output Shape
f	$8 \times 40 \times 512$	1×1 conv.	$8 \times 40 \times 448$
z	$8 \times 40 \times 64$	concat.	$8 \times 40 \times 512$
	$8 \times 40 \times 512$	5×5 deconv.	$16 \times 80 \times 256$
	$16 \times 80 \times 256$	5×5 deconv.	$32 \times 160 \times 128$
	$32 \times 160 \times 128$	5×5 deconv.	$64 \times 320 \times 64$
	$64 \times 320 \times 64$	1×1 conv.	$64 \times 320 \times 32$
	$64 \times 320 \times 32$	1×1 conv.	$64 \times 320 \times 3$

Table 2. Deep energy network architecture. All 3×3 convolutions use a stride of 2. $I_f = G(f, z)$, where G is the deep generator and f, z are its parameters.

Input	Input Shape	Operation	Output Shape
I_f	$64 \times 320 \times 3$	3×3 conv.	$32 \times 160 \times 64$
	$32 \times 160 \times 64$	3×3 conv.	$16 \times 80 \times 128$
	$16 \times 80 \times 128$	3×3 conv.	$8 \times 40 \times 256$
f	$8 \times 40 \times 512$	concat.	$8 \times 40 \times 768$
	$8 \times 40 \times 768$	1×1 conv.	$8 \times 40 \times 32$
	$8 \times 40 \times 32$	1×1 conv.	$8 \times 40 \times 3$
	$8 \times 40 \times 3$	energy term	1×1

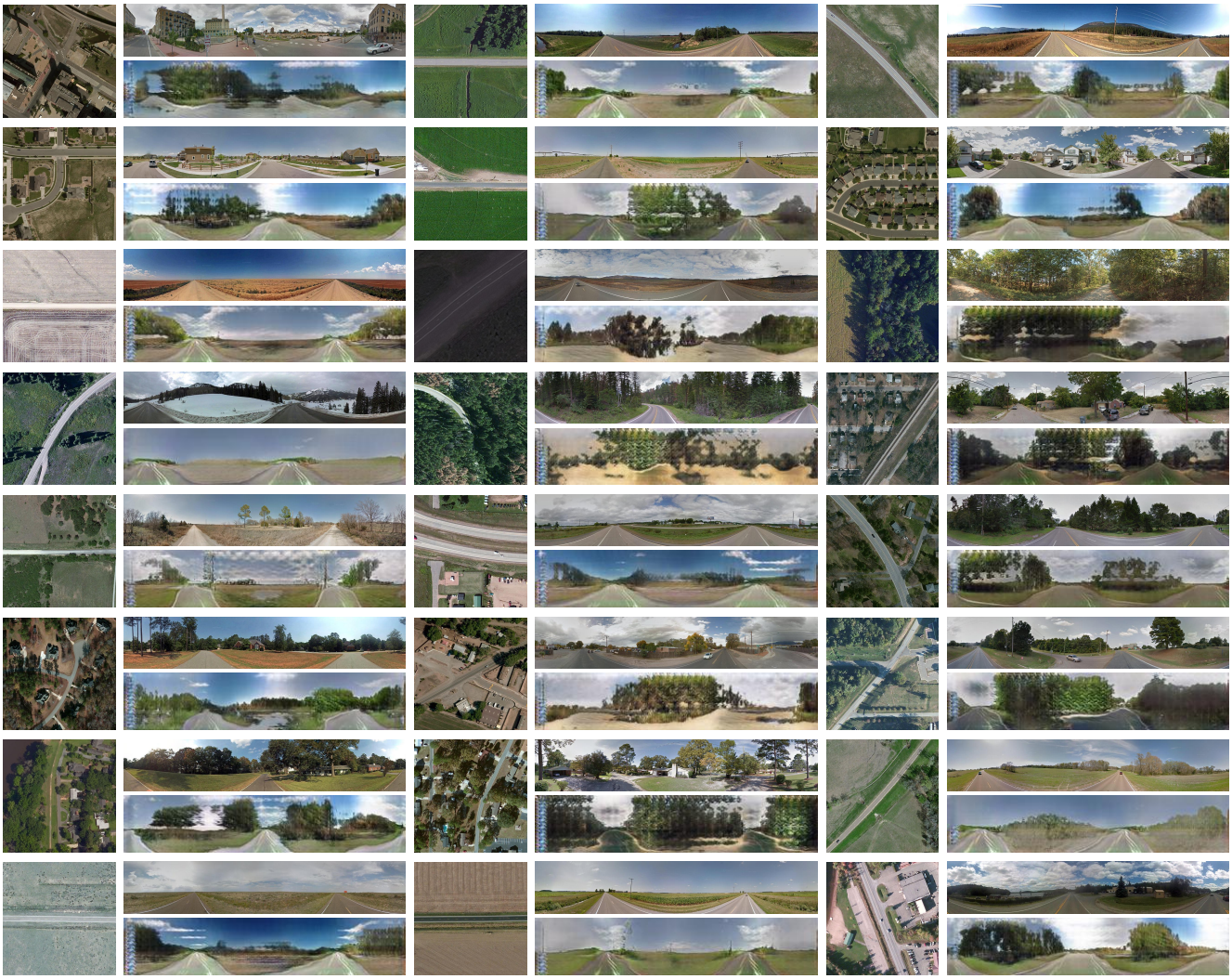


Figure 3. Randomly sampled test image results for synthesizing ground-level views. Each row shows an aerial image (left), its corresponding ground-level panorama (top-right), and predicted ground-level panorama (bottom-right).