Synthetic Generation of Satellite Images with Roads using ControlNet for Stable Diffusion

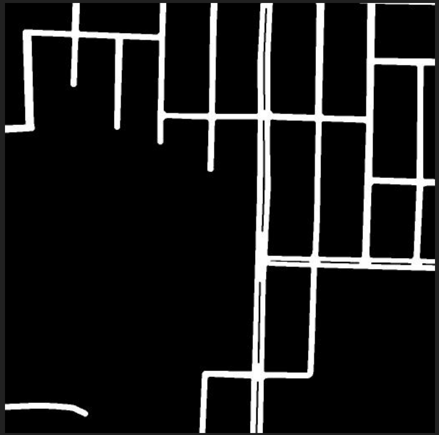
6.8300 Final Project Presentation

Sumedh Shenoy and Ram Goel

Problem Formulation

- **Our problem:** Creating *synthetic satellite images* based on (1) road masks and (2) text description of target image
- *Road mask* is black image with white lines for curves of road
- Goals for generated images:
    - Does not *add additional* roads, nor *delete* roads in our mask
    - *Natural looking* features surrounding roads (e.g. mountain ranges, forest areas, water bodies, etc.)

"aerial view of a rustic village"



Input Prompt *Input Mask* *Output*

Previous Approaches: Stable Diffusion with Inpainting

- Given mask and image, uses Stable Diffusion to **replace** non-road parts of image

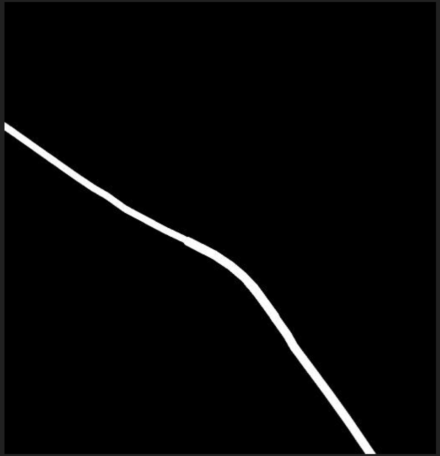
Left: *Inputted* Road Mask, Middle: *Inputted* Full Image, Right: *Outputted* Stable Diffusion Generated Image

Limitation: Inpainting needs mask along with full image for each image generated, so cannot synthesize new images from only a road mask.

Turn to ControlNet!

ControlNet provides a backbone to Stable Diffusion, giving the **ability to condition** Stable Diffusion on more than just text prompt.

- Inference: Takes in mask and text prompt, outputs **generated image**
- We use road masks as *input condition*; this is what guides diffusion spatially beyond text
- Key advantage from inpainting: at generation time, only requires a road mask, not full image



Input: road mask, and prompt: “aerial view of a road by water, high-quality, 8K”

Output: Generated image

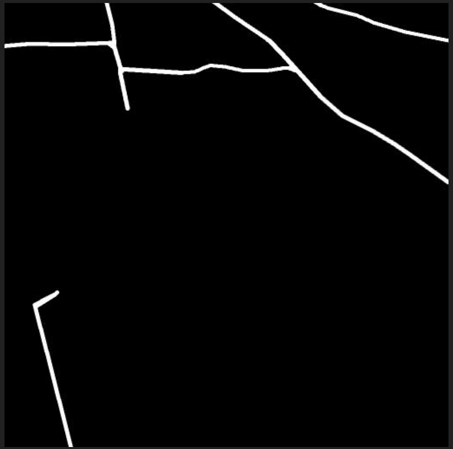
ControlNet Training Procedure

- Training data composed of tuples of:
  - Road mask condition
  - Associated image with road mask
  - Associated prompt with image
- Prompt is replaced with empty string “” with 50% probability, to learn more semantics from mask

"aerial view of a small village in a forested area"

Data Collection

- We used a dataset curated by DeepGlobe 2018 Challenge
  - Over 6000 pairs of road masks & images



Road Mask

Full Image

- We generated captions for full images by feeding into BLIP (image captioning model)
  - Example: “aerial view of a desert area with a small town in the middle”

Data Modifications

- When rescaled in size, roads became extremely thin
  - Potentially not strong enough guidance as controls
- New dataset: modify road masks by **thickening the roads** via dilation

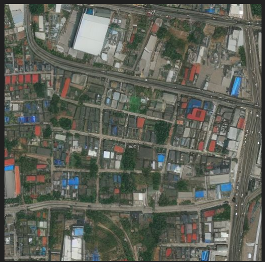
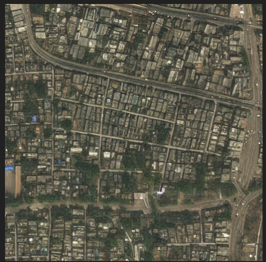
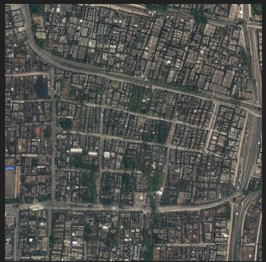
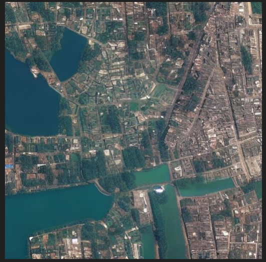
Unmodified Road Mask

Modified Road Mask

Our 3 Models

	Training Procedure	Training Dataset
Model 1	Standard ControlNet	DeepGlobe + BLIP
Model 2	Standard ControlNet	Thickened DeepGlobe + BLIP
Model 3	Elevated Probability of Empty Prompt	DeepGlobe + BLIP

Results: Dense Road Maps Synthesize Well!

Prompt	Original Mask	Model 1: trained on unmodified masks	Model 2: trained on dilated masks	Model 3: Trained with fewer prompts
“overhead satellite image, busy city, extremely good quality, road, high resolution”				
“aerial view of a barren desert”				
“overhead satellite image, busy city with a lake, extremely good quality”				

Results: Good Performance for Sparse Maps

Prompt

Original Mask

Model 1: trained on unmodified masks

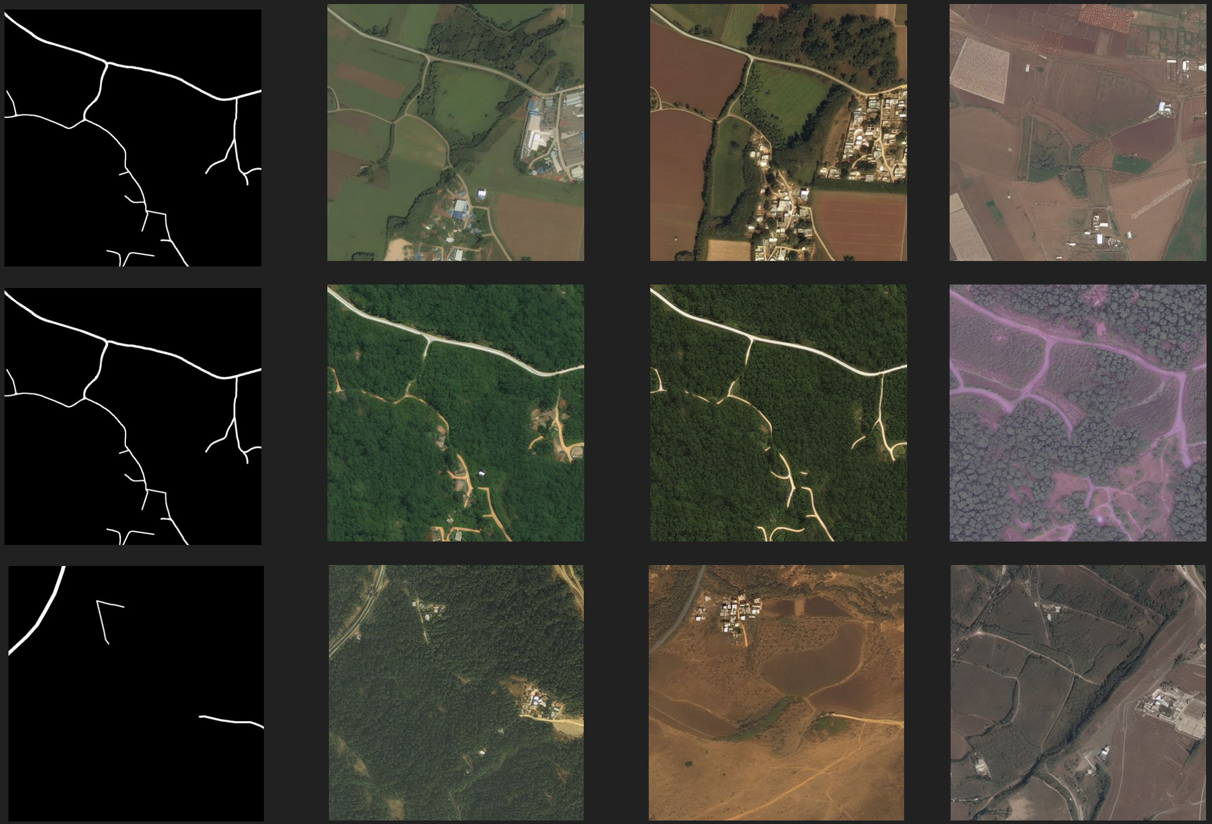
Model 2: trained on dilated masks

Model 3: Trained with fewer prompts

“aerial view of a farm”

“detailed aerial view of a forest”

“aerial view of a mountainous area with a road”



Results (cont.): Model 3 best prompt fidelity

Original Mask

Model 1: trained on unmodified masks

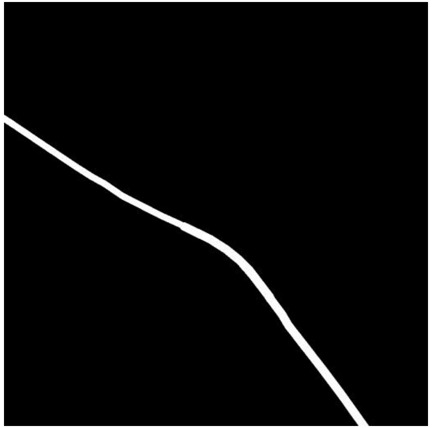
Model 2: trained on dilated masks

Model 3: Trained with fewer given prompts

Prompt: “overhead aerial photograph, **cloudy day**, landscape with a river running through it, extremely detailed”

Limitation: Model 1 and Model 2 exhibit overfitting to training prompts

- “Cloud” sparse in prompt dataset

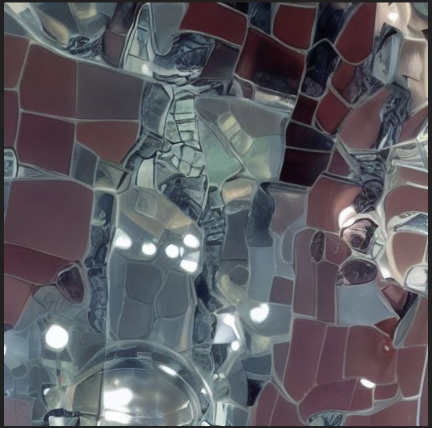



Prompt: “overhead aerial photograph, **cloudy day**, landscape with a river running through it, extremely detailed”

Results: Model 3 Best for Sparse Prompts

Original Mask

Model 1: trained on unmodified masks



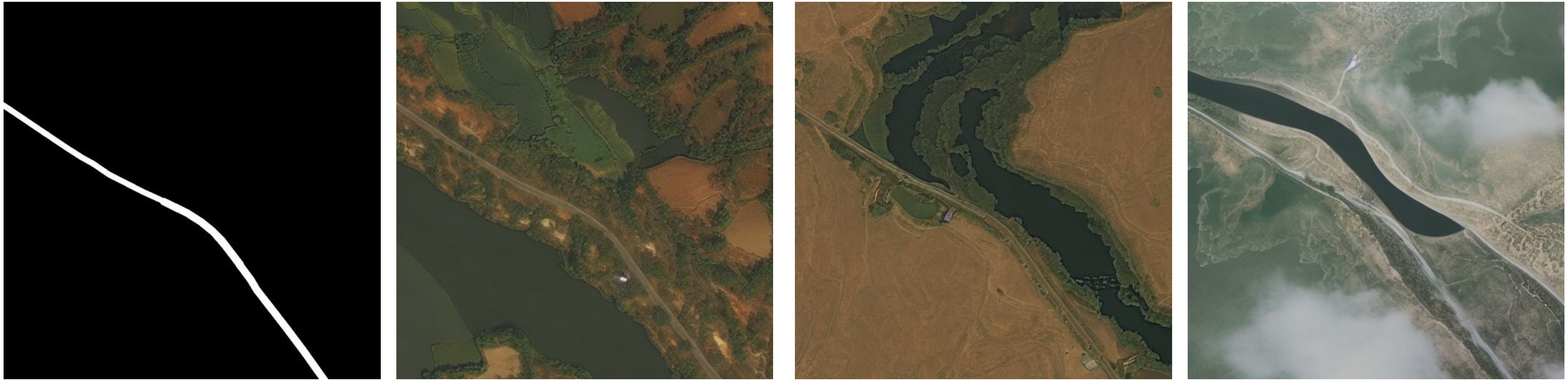
Model 2: trained on dilated masks

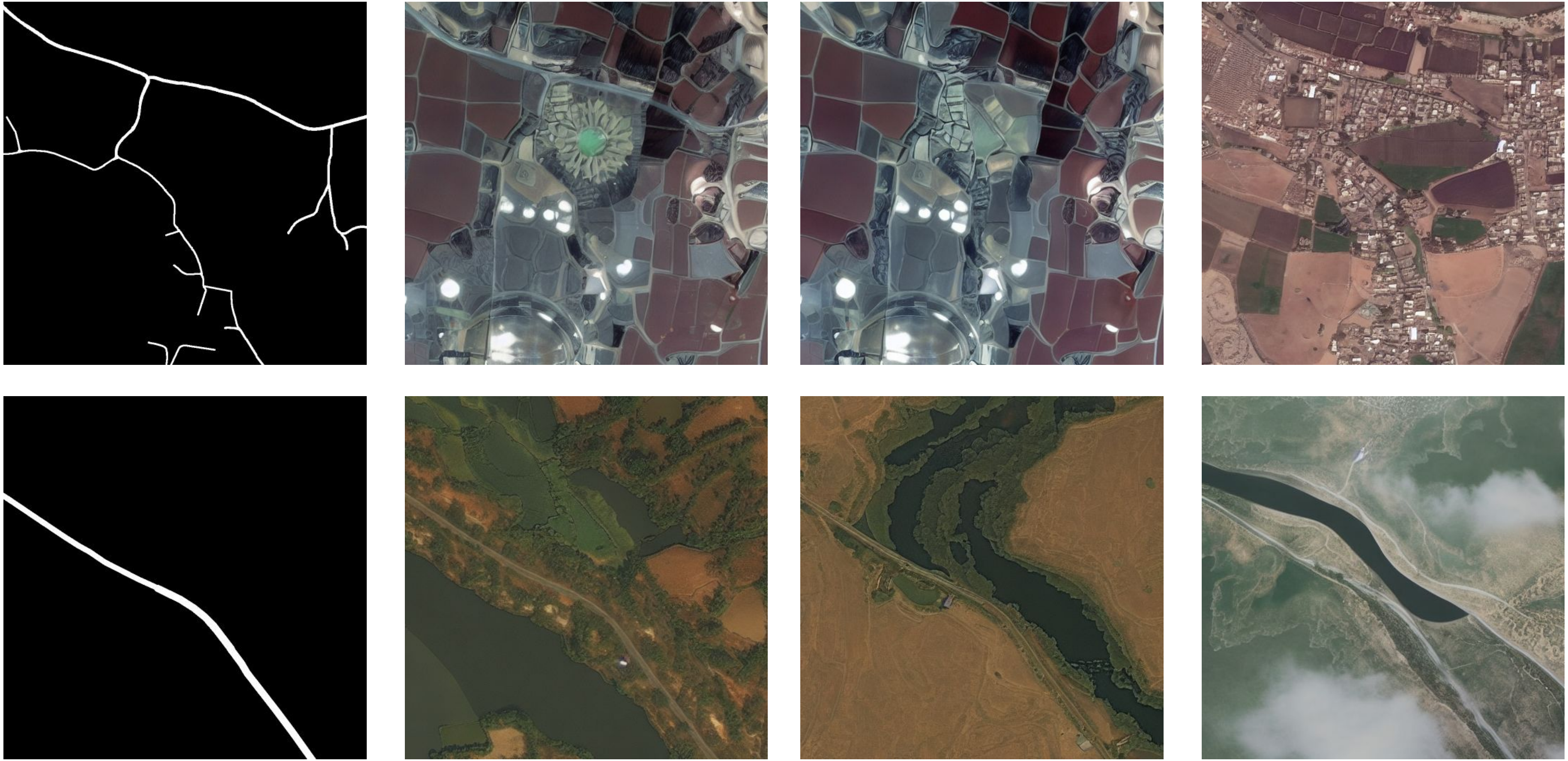
Model 3: Trained with fewer given prompts

Prompt: “”

Limitation: Model 1 and Model 2 exhibit overdependence to prompt

- Model 3 performs better due to fewer supplied prompts



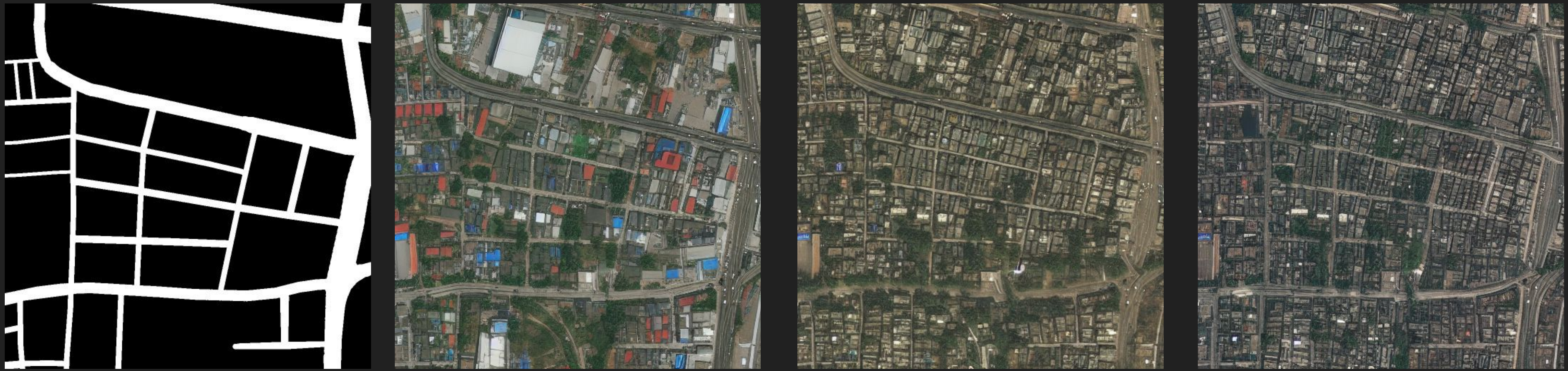


Results: Summarized

- Model 3 performs the best. Better prompt fidelity (as seen with clouds, e.g.)
- Models 1 and 2 do not handle prompts not well-represented in the training set well
- All models perform well on both sparse and dense road maps
- Limitations:
  - Training prompt vocabulary is limited
- Further tests:
  - Try more types of terrain
  - Greater vocabulary
  - Find optimal prompt dropout ratio

Thank you!

Results (cont.)



Original Mask

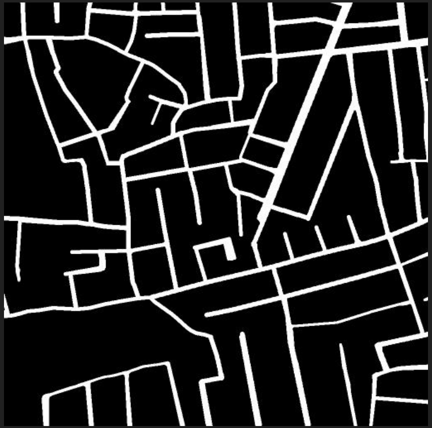
Model 1: trained on unmodified masks

Model 2: trained on dilated masks

Model 3: Trained with fewer given prompts

Prompt: “overhead satellite image, busy city, extremely good quality, road, high resolution”

Results (cont.)



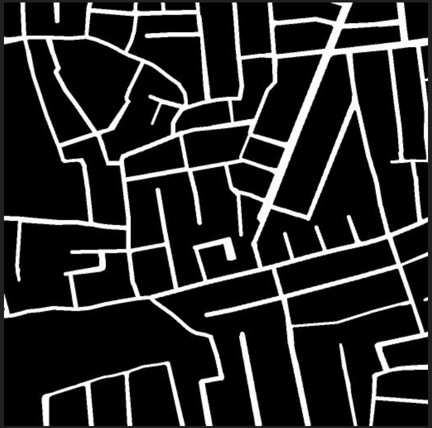
Original Mask

Model 1: trained on unmodified masks

Model 2: trained on dilated masks

Model 3: Trained with fewer given prompts

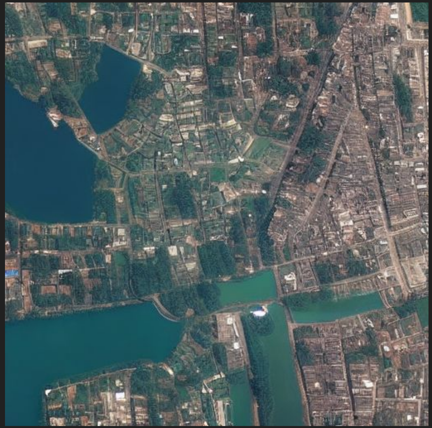
Prompt: “overhead satellite image of a busy city with a lake with clouds”



Original Mask

Base ControlNet model we trained

Trained on condition of dilated road masks



Trained with fewer given prompts

Prompt: “overhead satellite image, busy city with a lake, extremely good quality”

Results (cont.)

Original Mask

Model 1: trained on unmodified masks

Model 2: trained on dilated masks

Model 3: Trained with fewer given prompts

Prompt: “aerial view of a barren desert”

Results (cont.)

Original Mask

Model 1: trained on unmodified masks

Model 2: trained on dilated masks

Model 3: Trained with fewer given prompts

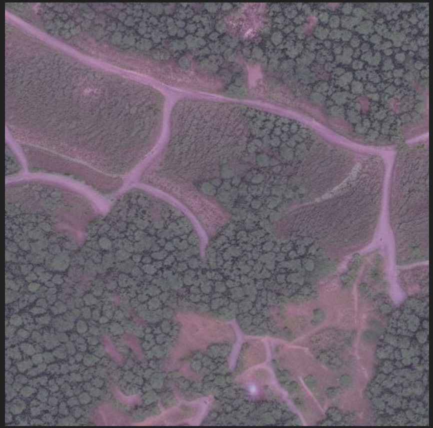
Prompt: “aerial view of a farm”

Results (cont.)

Original Mask

Model 1: trained on unmodified masks

Model 2: trained on dilated masks



Model 3: Trained with fewer given prompts

Prompt: “detailed aerial view of a forest”

Results (cont.)

Original Mask

Model 1: trained on unmodified masks

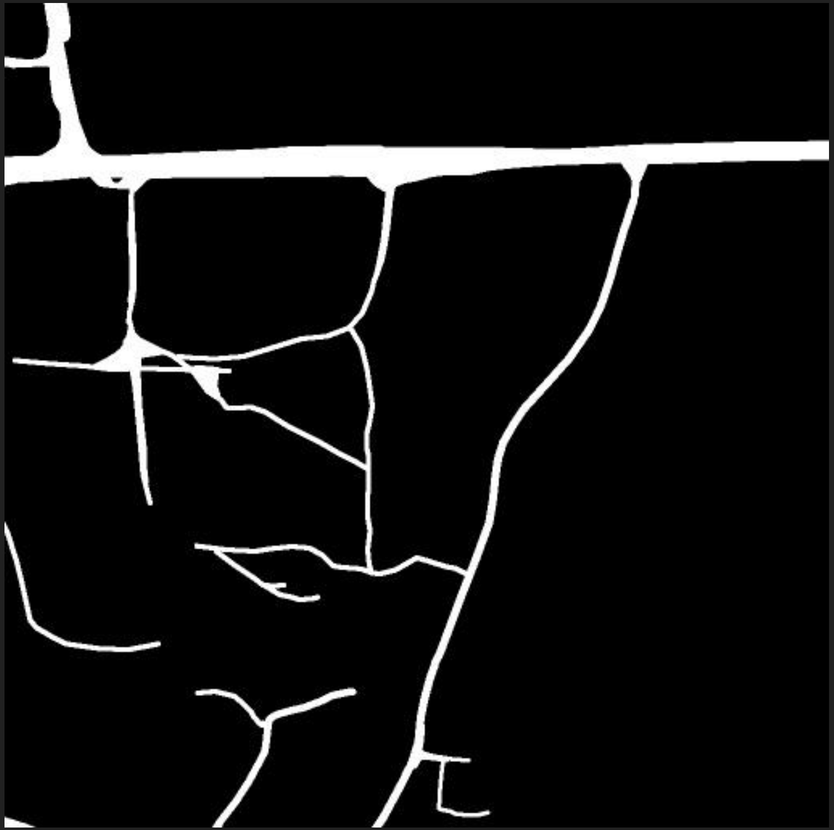
Model 2: trained on dilated masks

Model 3: Trained with fewer given prompts

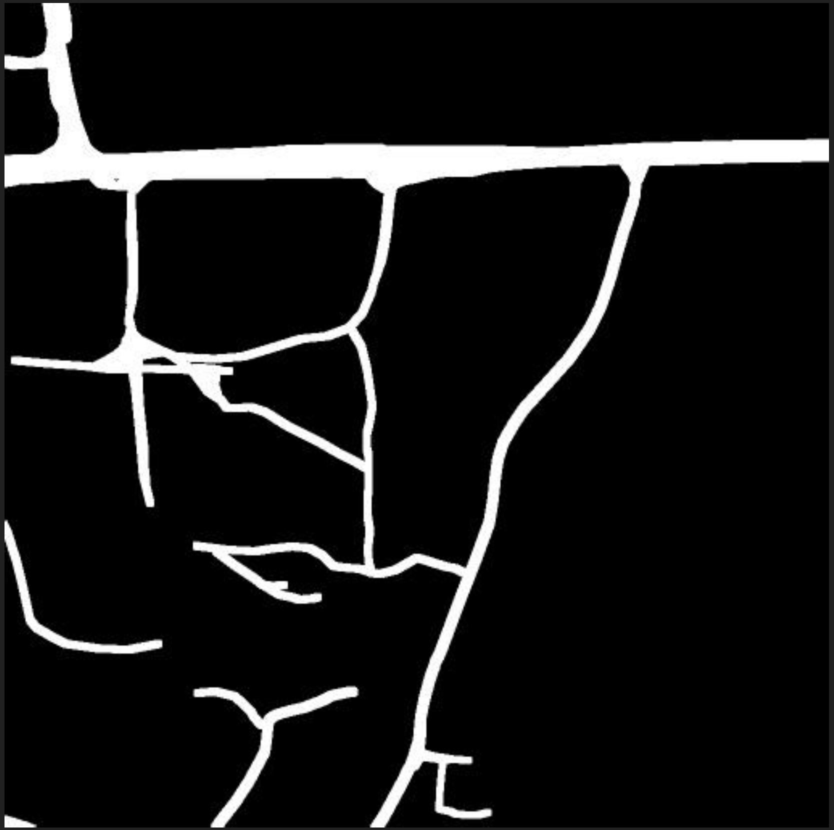
Prompt: “aerial view of a mountainous area with a road”

Unmodified road mask

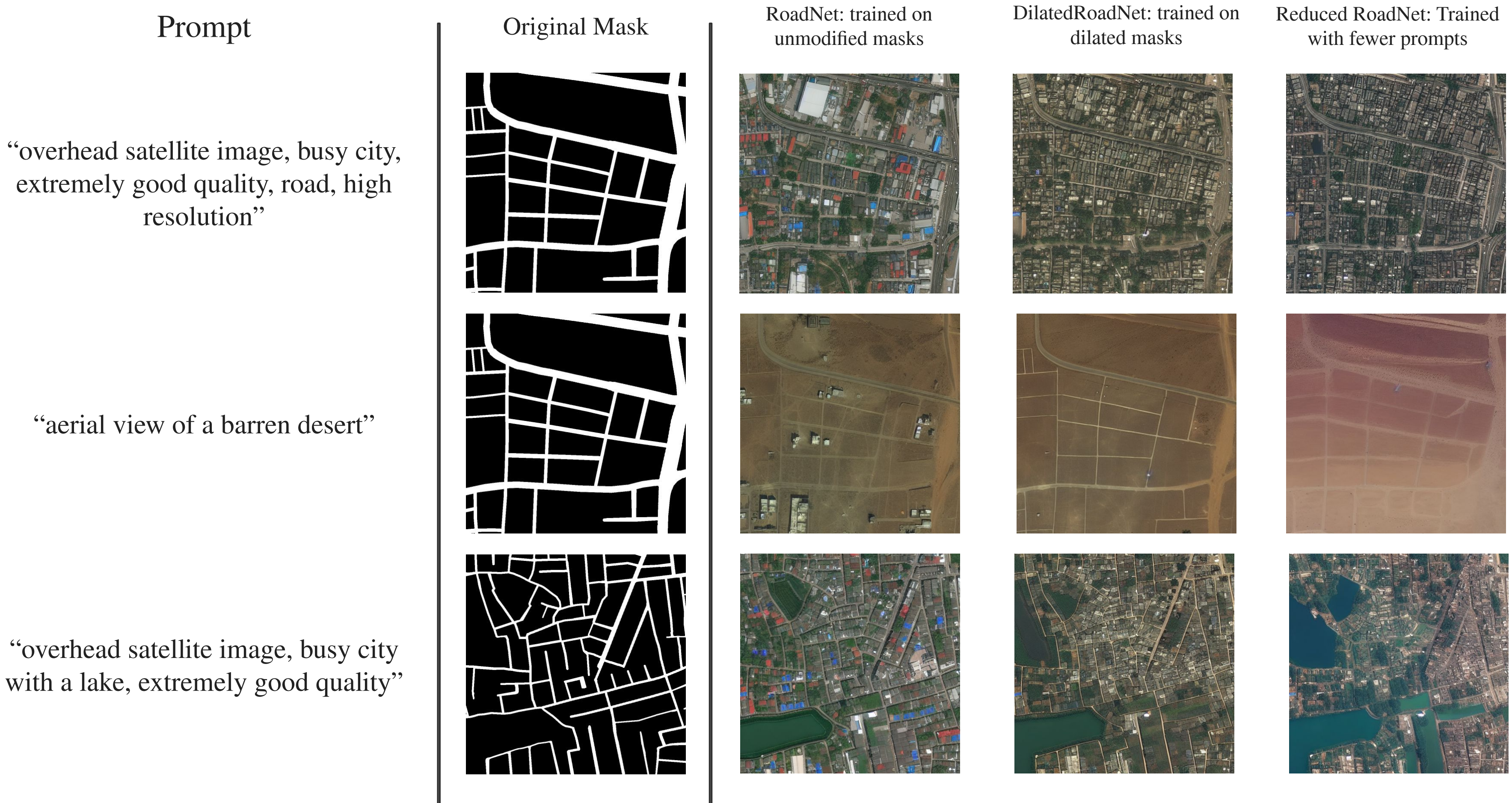
Modified road mask



Unmodified road mask



Modified road mask


Prompt
Original Mask
RoadNet: trained on unmodified masks
DilatedRoadNet: trained on dilated masks
Reduced RoadNet: Trained with fewer prompts
“overhead satellite image, busy city, extremely good quality, road, high resolution”
“aerial view of a barren desert”
“overhead satellite image, busy city with a lake, extremely good quality”

Results: Good Performance for Sparse Maps

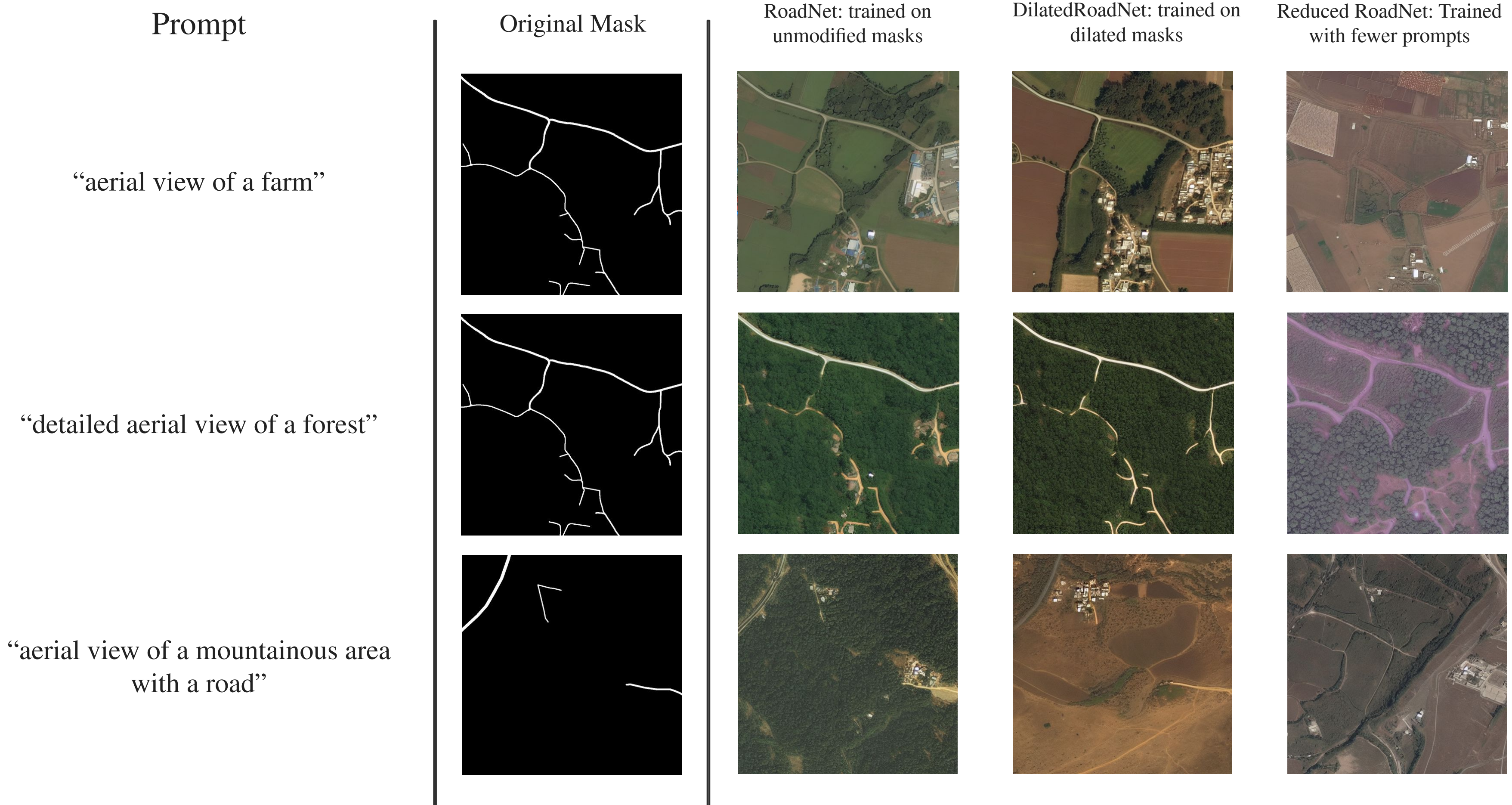
Model 2: trained on dilated masks

Model 3: Trained with fewer prompts

“aerial view of a farm”

“detailed aerial view of a forest”

“aerial view of a mountainous area with a road”


Prompt
Original Mask
RoadNet: trained on unmodified masks
DilatedRoadNet: trained on dilated masks
Reduced RoadNet: Trained with fewer prompts
“aerial view of a farm”
“detailed aerial view of a forest”
“aerial view of a mountainous area with a road”