

Learning by Looking

Solving Formal Problems from Images

Presented by: Lance Bassett, Jiahang Li

Humans solve “algorithmic” problems visually

- Correct solutions “look” correct
- Solve by looking, very little formal process

KIDS MENU
All meals served with choice of fries, side salad, seasonal vegetables or apple sauce.

Cheese Burger
Kids size with American cheese and a pickle. \$5

Chicken Tenders
Three breaded tenders served with choice of sauce.

Grilled Chicken Breast
\$7

Hot Ham & Cheese
All natural ham, American cheese & honey mustard with your choice of wheat or white bread. \$5

Grilled Cheese
Pancake pressed with your choice of white or wheat bread. \$4

Mac n' Cheese
Creamy mac made with full noodles. \$5

Butter Noodles
Fried, butter and parmesan. \$4

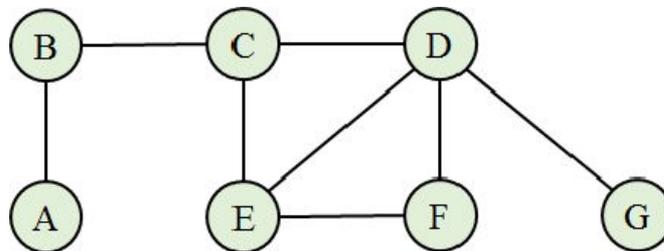
Crab Cake Slider
Mini jumbo lump crab cake on a bun served with a pickle. \$5

Help the puppy find his owner!

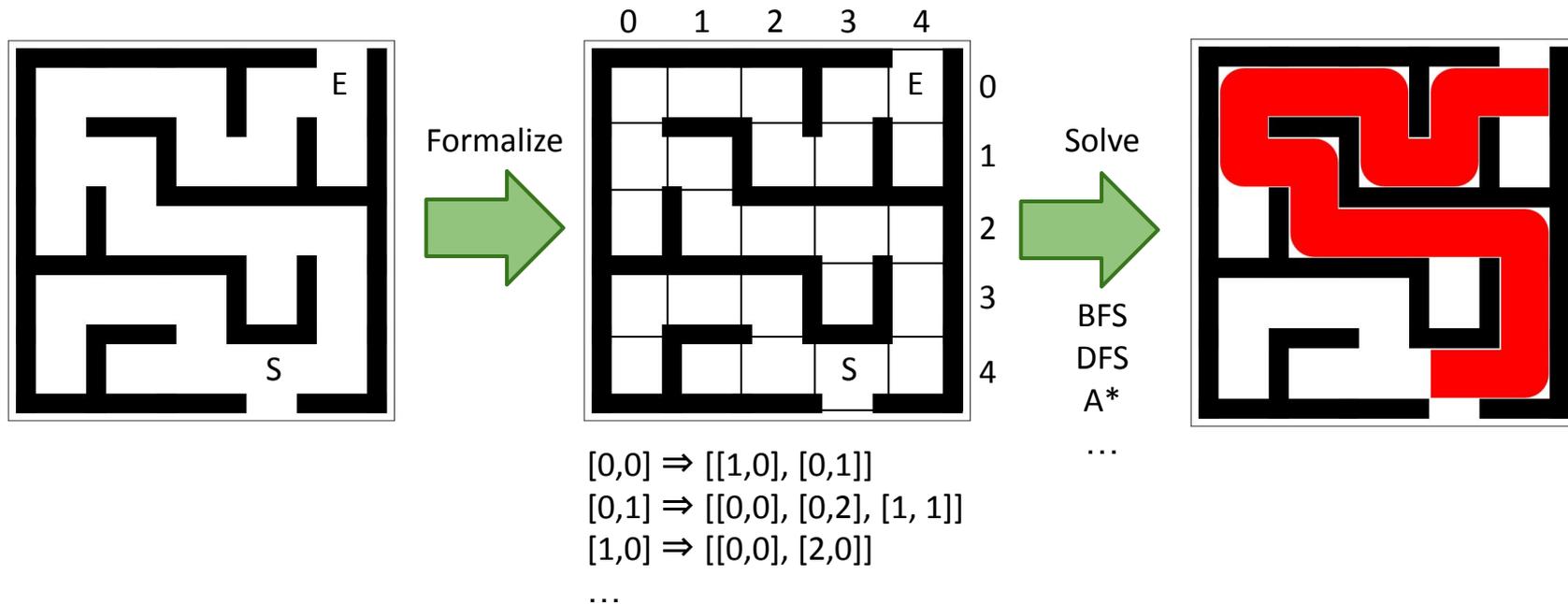
DOT GAME: The objective is to create a face by making a line from one dot to the other. You can only make one line per turn. The person with the most faces at the end of the game wins!

TIC TAC TOE

Dan's Diner
3 South Main Street | Boonshoro MD 21713
301-432-9224 | www.DRINTH.com



Can we obtain visual problem solving through neural networks?



Can we obtain visual problem solving through neural networks?

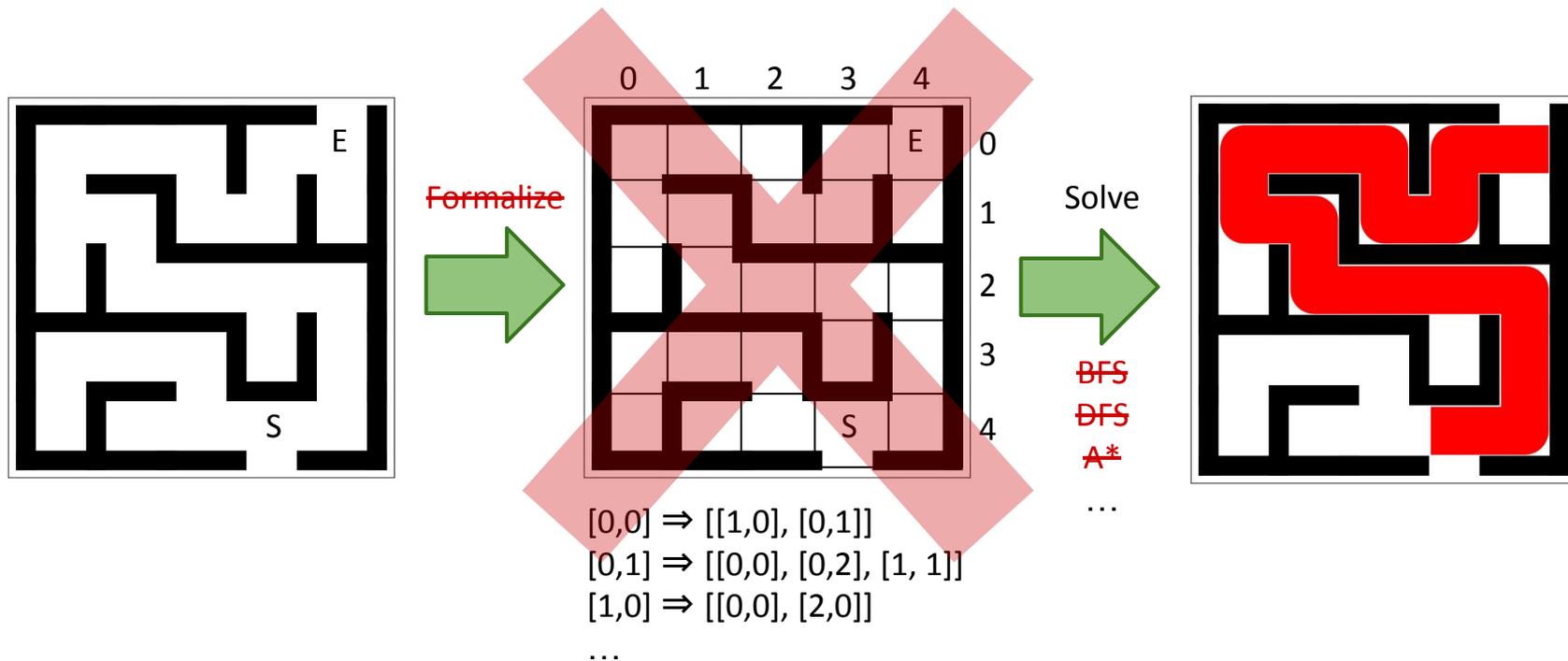


Image to Image Translation with pix2pix

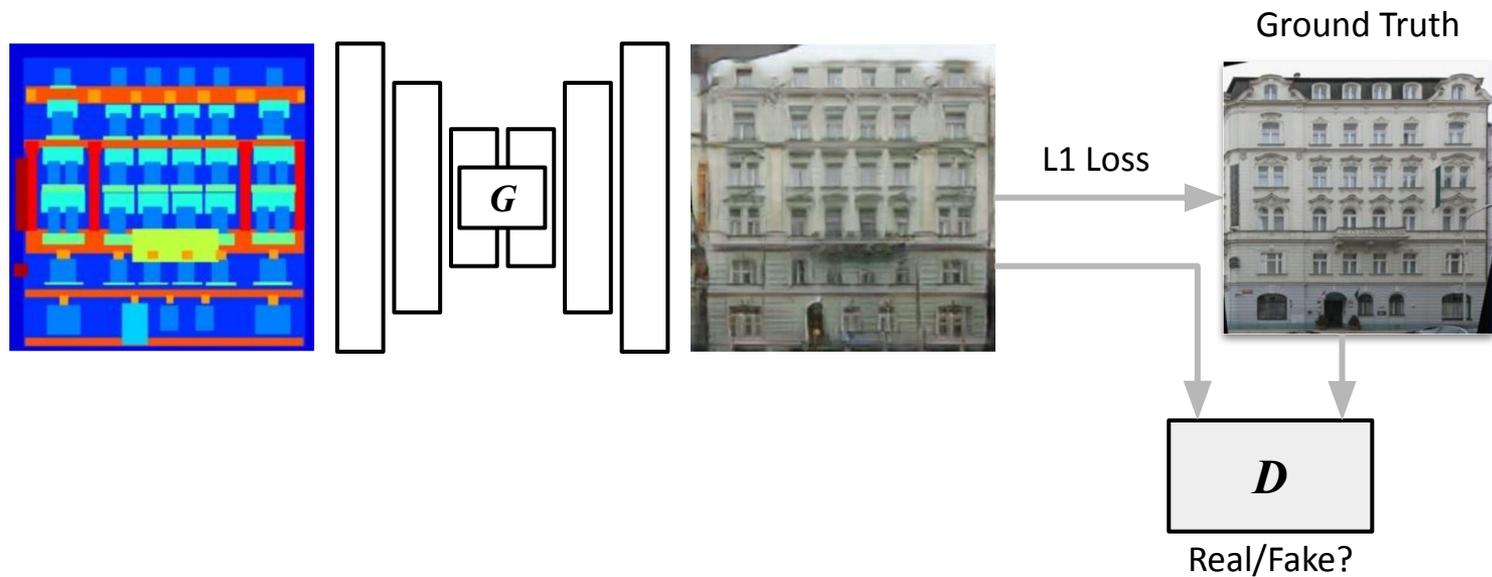
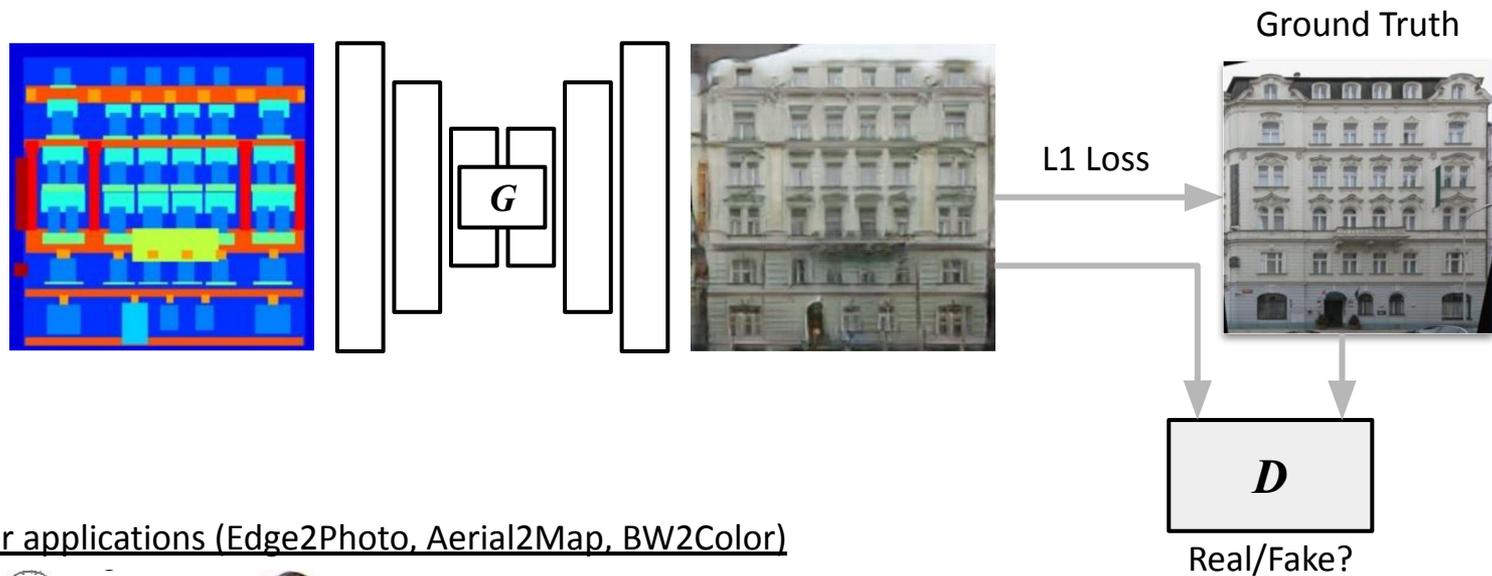


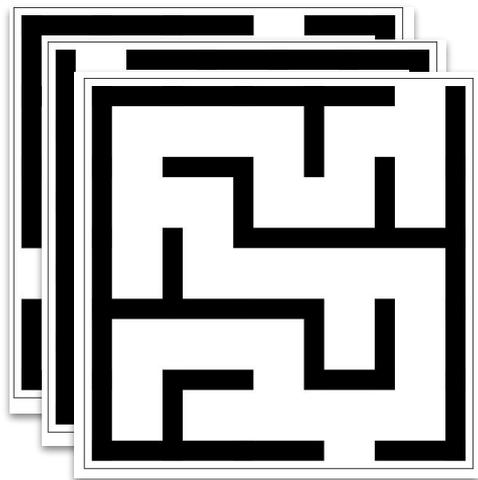
Image to Image Translation with pix2pix



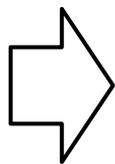
Other applications (Edge2Photo, Aerial2Map, BW2Color)



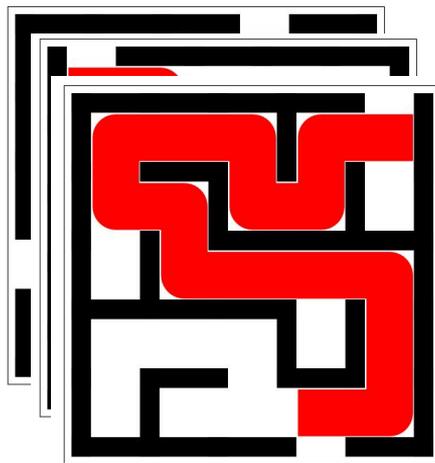
Application to mazes



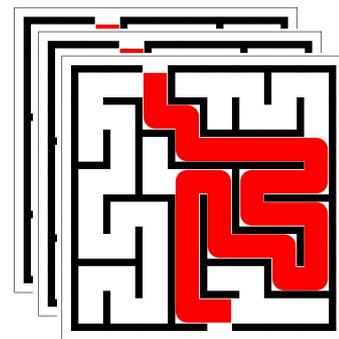
Source



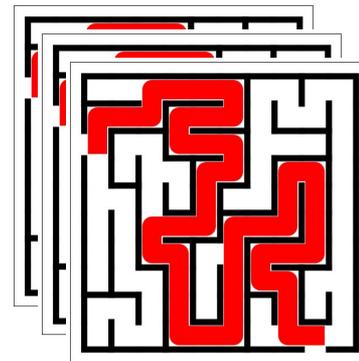
(5x5)



Target



(8x8)

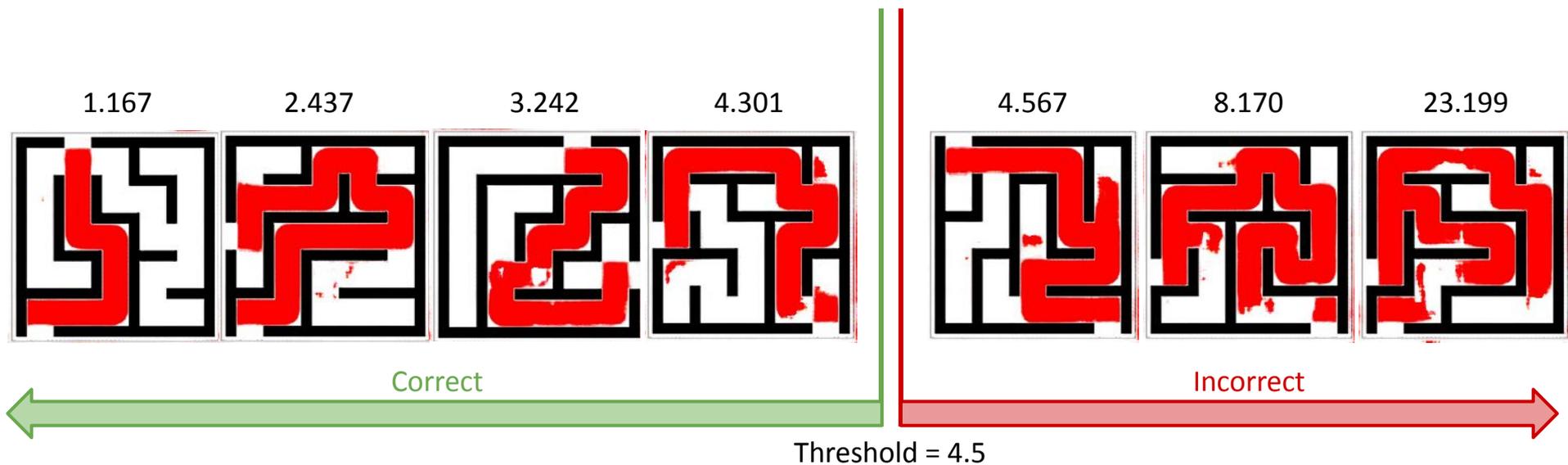


(10x10)

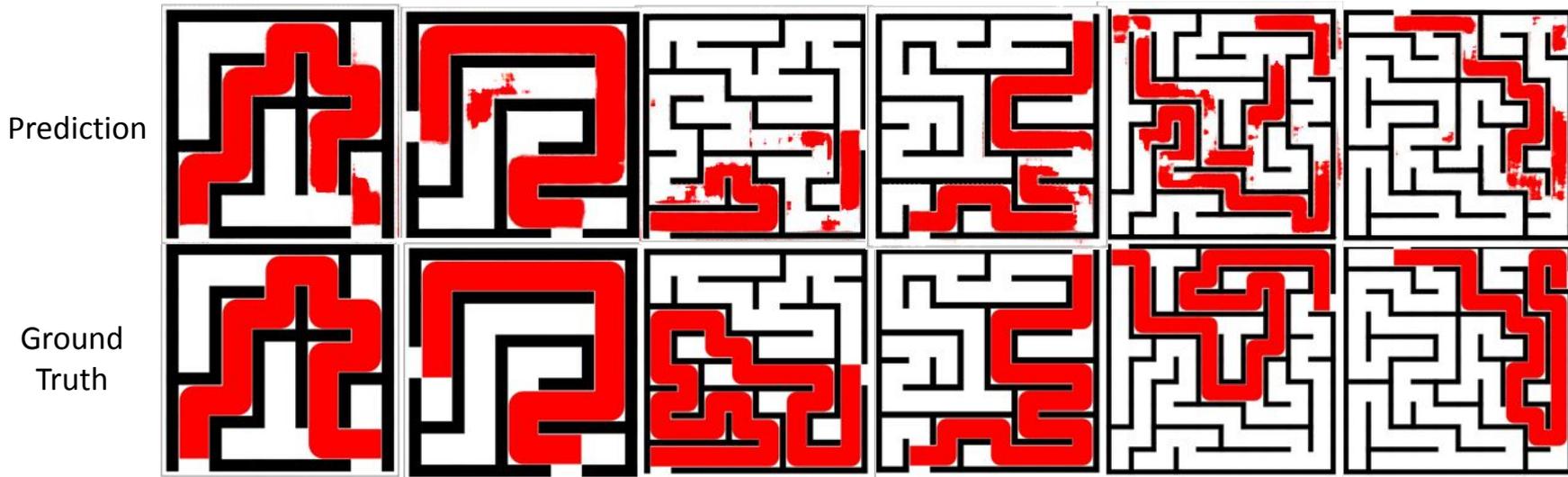
4000 train, 1000 test for all sizes

L1 loss metric

Correctness is visually estimated based on L1 loss between generated and ground truth



Initial results

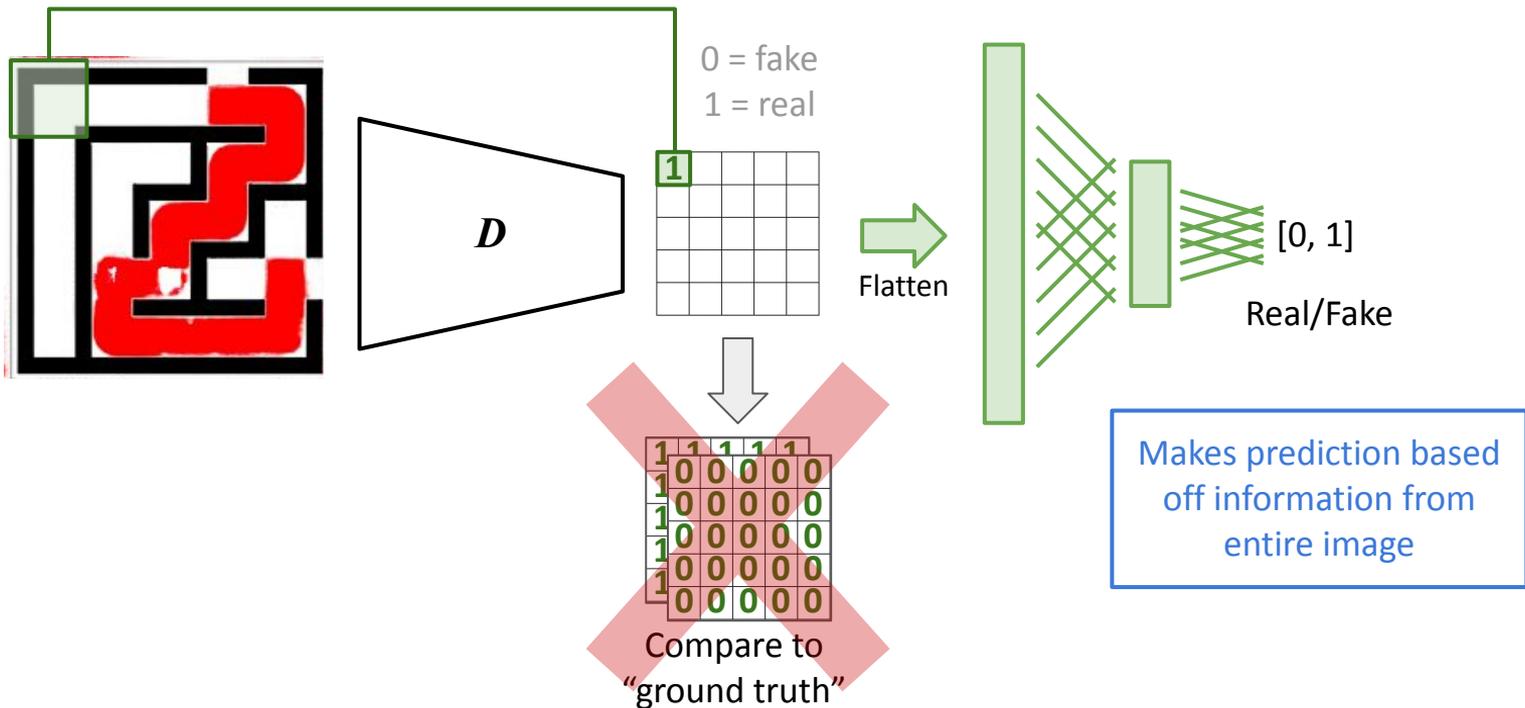


| Maze Size | L1 Accuracy | Mean L1 Loss | Median L1 Loss |
|------------------|--------------------|---------------------|-----------------------|
| 5x5 | .796 | 3.558 | 1.874 |
| 8x8 | .599 | 5.734 | 3.154 |
| 10x10 | .259 | 9.989 | 8.769 |

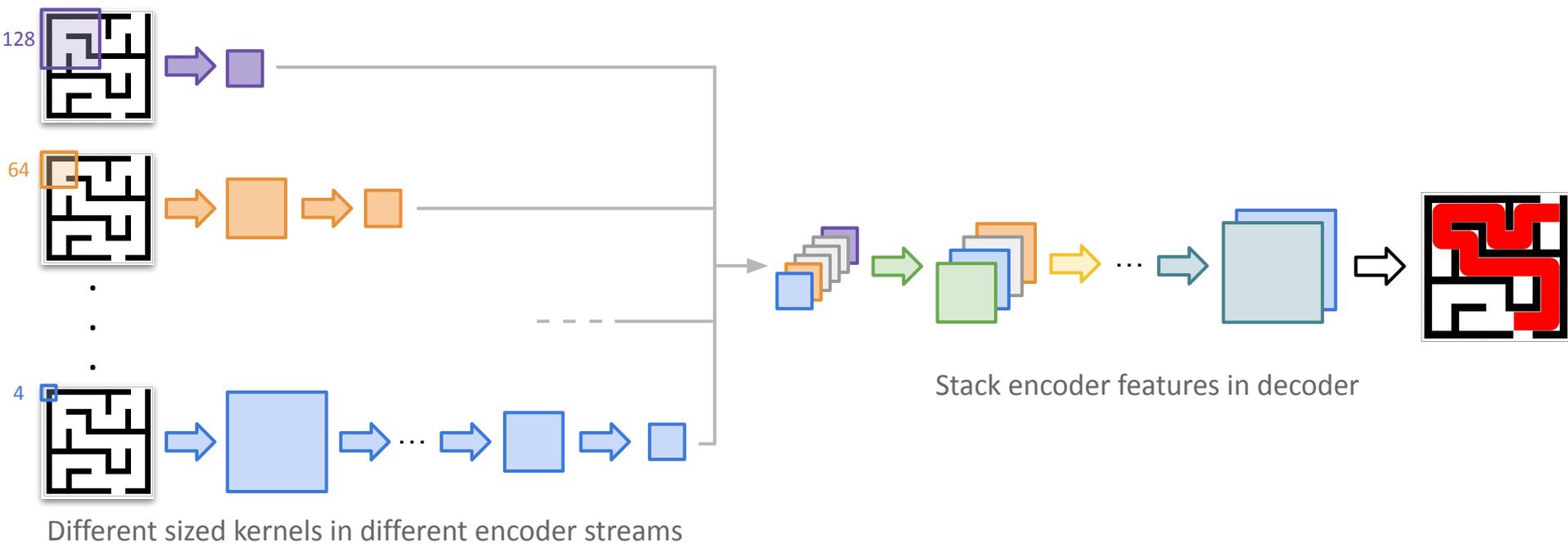
Architectural Solutions: Non-patch based discriminator

Discriminator considers local patches

Assumes independence between pixels separated by > patch diameter



Architectural Solutions: Multiscale Generator

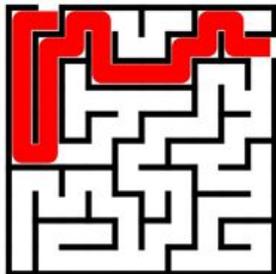
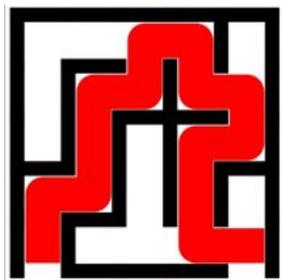


Quantitative Results across Models

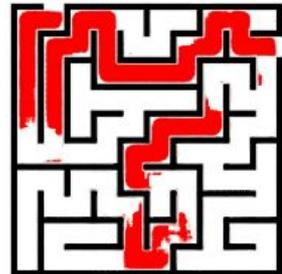
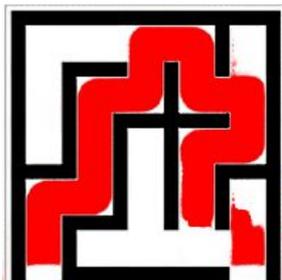
| Size | Model | L1 Accuracy | Mean L1 Loss | Median L1 Loss |
|-------|---------------------|-------------|--------------|----------------|
| 5x5 | Baseline | .796 | 3.558 | 1.874 |
| | Global Disc. | .814 | 3.139 | 1.723 |
| | Multiscale Gen. | .452 | 7.526 | 5.101 |
| 8x8 | Baseline | .599 | 5.734 | 3.154 |
| | Global Disc. | .649 | 5.298 | 2.683 |
| | Multiscale Gen. | .123 | 12.764 | 12.988 |
| 10x10 | Baseline | .259 | 9.989 | 8.769 |
| | Global Disc. | .386 | 8.653 | 6.440 |
| | Multiscale Gen. | .034 | 18.224 | 17.167 |

Qualitative Results Across Models

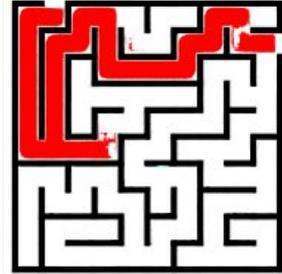
Ground Truth



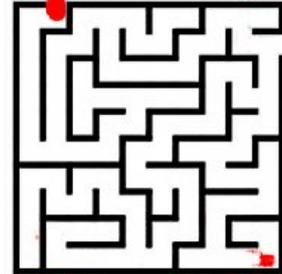
Baseline



Global Disc.



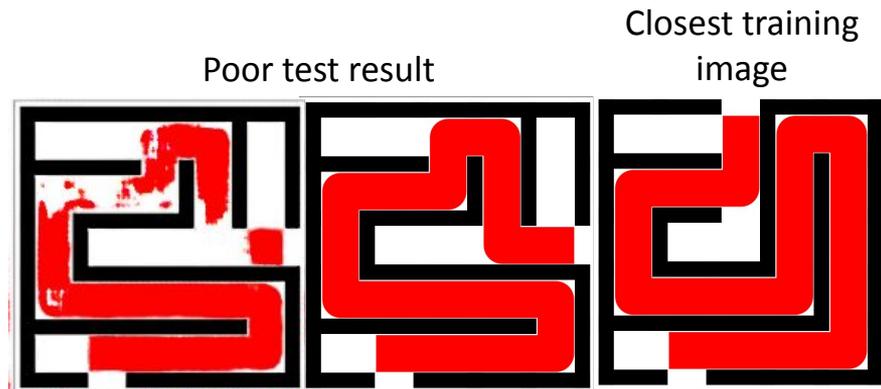
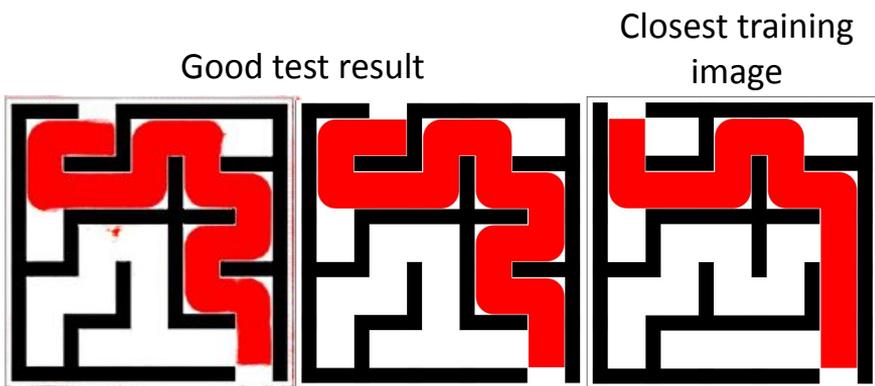
Multiscale Gen.



Discussion

Generalization in larger mazes

- 5x5 inherently less available variation in data
- Could find success with more data or alternate global architecture



Mystery: Mistake occurs in similar region

Future Considerations

- Transformer: Attention could learn globally for each pixel
- pix2pixHD: Could help as problem scales visually
- Diffusion model (ControlNet?): Use generative prior
- More data: Would help for larger mazes (wary of enumerating all possibilities)

Thank you!

