Single-Image Stereo Depth Estimation using GANs

CS231A - Winter 2021
Sharan Ramjee, Nikhil Parab, Naveed Zaman

Stanford University

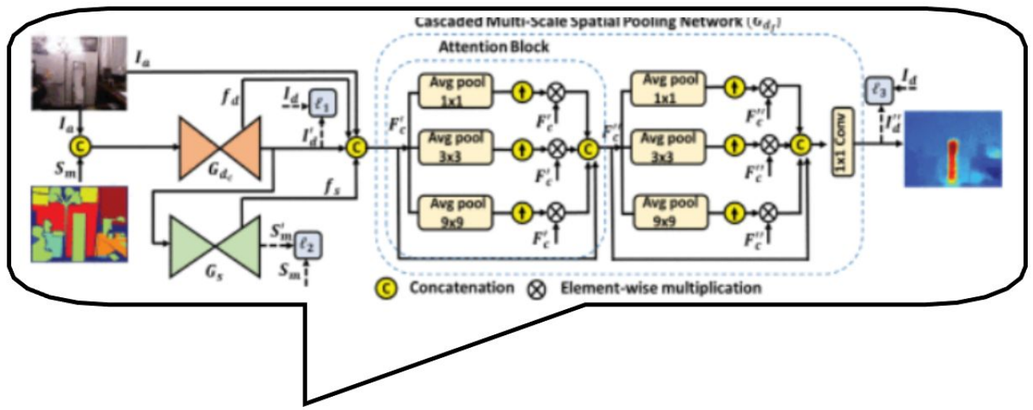
Introduction

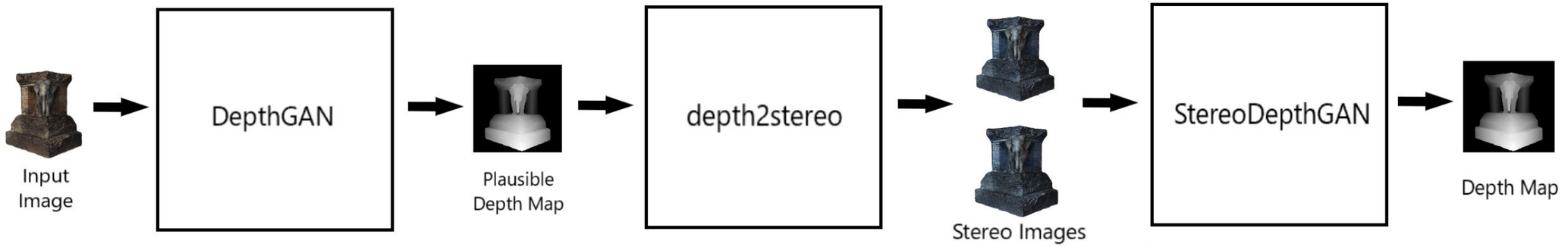
- Motivation:
  - Traditional depth estimation methods require multiple viewpoint images
  - Deep Neural Network (DNN) based monocular methods look promising

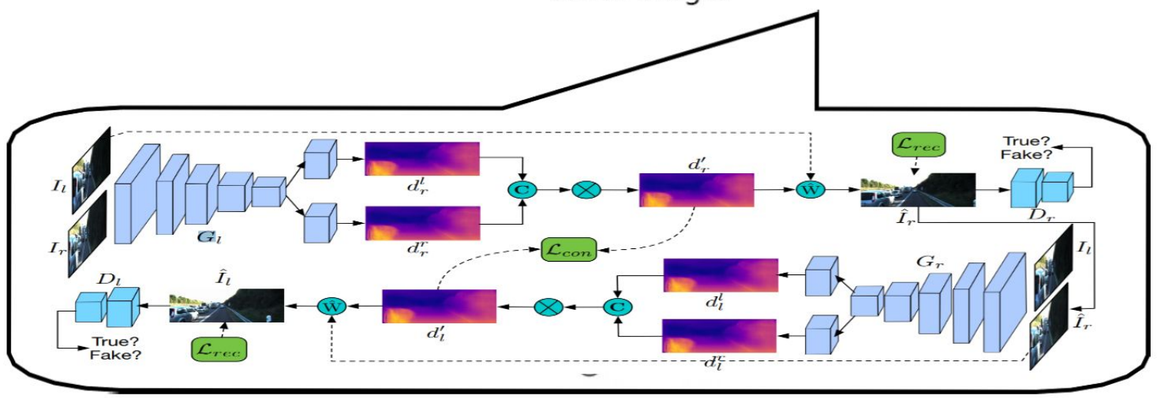
- Challenge:
  - Stereo DNN methods still superior to monocular DNN techniques
  - Stereo depth estimation requires complex setup (infeasible)

- Our Approach: A novel monocular depth estimation pipeline using GAN-based stereo depth estimation methods:
  - **DepthGAN**: Left image → ”plausible” depth map
  - **Depth2Stereo**: Left image + “plausible” depth map → Right image
  - **StereoDepthGAN**: Left image + Right image → Final depth map

Single-Image Stereo Depth Estimation Pipeline







DepthGAN

- Overview: Re-formulate depth estimation task as image reconstruction task using the reconstruction-based GAN architecture

- Input: Left view image
- Output: “Plausible” depth map (left-to-right disparity map)

- Generator: Generates the depth map using the single left view image and uses the warping function f_w to reconstruct the original left view image
  - **L1 reconstruction loss**: Minimize absolute per-pixel distance
  - **Structural similarity (SSIM) loss**: Measure perceived quality
  - **Consistency loss**: Enforce consistency between disparity maps
  - **Disparity smoothness loss**: Enforce smooth disparities

- Discriminator: Trained to discern between real and generated images
  - **Cross-entropy loss**: Vanilla GAN discriminator loss

Depth2Stereo Algorithm

- Overview: Objects popping out of the image in the depth map are farther apart from each other in comparison to objects sinking into the image

- Input: Left view image and “plausible” depth map (from DepthGAN)
- Output: Right view image

- Two step process:
  - **Tear identification step**: Tears occur along regions where depth map intensities increase over entire pixel range along the row. These ranges of pixels are torn (shifted to the right) and are assigned a value of 0
  - **Tear interpolation step**: The torn pixels (with value 0) are assigned either the average of the two neighboring pixels or the value of the left pixel (empirically better)

StereoDepthGAN

- Overview: Uses cycled generative adversarial networks to generate left-to-right disparity map (from original left image) and a right-to-left-disparity map (from depth2stereo generated right image), which are then enhanced using a 1 x 1 convolution layer

- Input: Left view image and right view image (from depth2stereo)
- Output: Final depth map

- Generator: Similar to DepthGAN, but reconstructs both left and right view images using corresponding depth maps generated
  - **L1 reconstruction loss**: Minimize absolute per-pixel distance
  - **Consistency loss**: Enforce consistency between disparity maps

- Discriminator: Trained to discern between real and generated images
  - **Cross-entropy loss**: Vanilla GAN discriminator loss

NYU Depth V2 Dataset

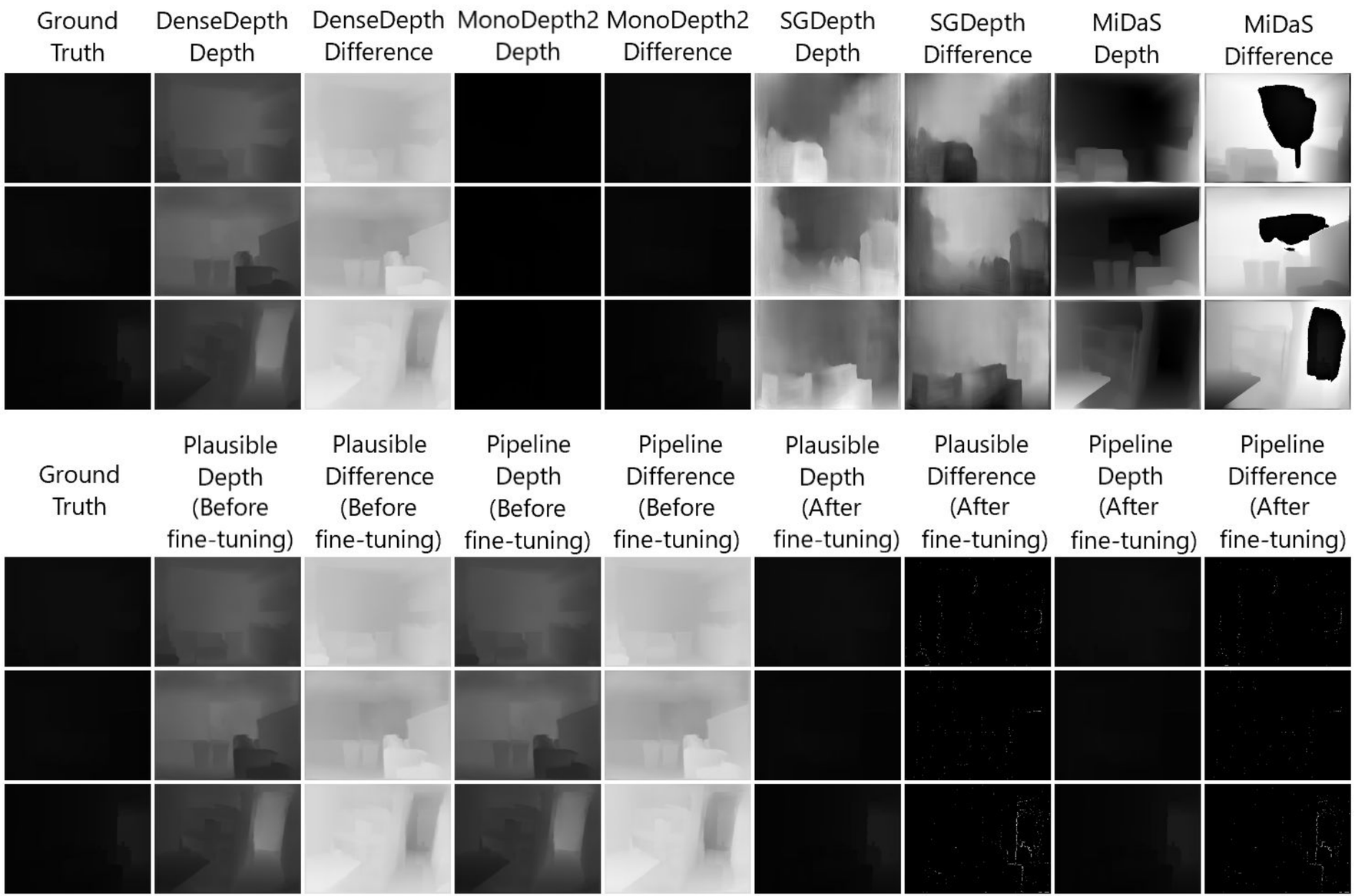
- NYU Depth V2 Dataset:
  ~120K rgb-depth image pairs captured using Microsoft Kinect
  resolution: 640x480, depth-map upper bound = 10m
  Image Inpainting for missing depth map
- Train Dataset:
  Subset of NYU Depth V2 Dataset containing 50,000 samples
- Test Dataset:
  Official NYU Depth V2 test dataset containing 654 samples

Quantitative Results

- Metrics: Absolute Relative Distance (ARD), Squared Relative Distance (SRD), Root Mean Square Error (RMSE), log Root Mean Square Error (log RMSE). Lower values are better across all metrics.

Method	ARD	SRD	RMSE	log RMSE
DenseDepth [1]	0.74	0.56	0.74	1.48
MonoDepth2 [7]	0.534	0.381	0.612	1.521
SGDepth [10]	0.550	0.345	0.584	1.027
MiDaS [16]	0.024	0.015	0.027	0.057
Ours - Plausible	0.792	0.646	0.802	2.560
Ours - Final	0.745	0.567	0.751	1.499
Ours - Plausible (After fine-tuning)	0.025	0.002	0.027	0.056
Ours - Final (After fine-tuning)	**0.019**	**0.013**	**0.022**	**0.046**

Qualitative Results



Summary

- Results:
  - Some baselines outperform pipeline generated “plausible” depth map
  - Whole pipeline outperforms all baselines considered

- Pipeline advantages:
  - Stereo depth estimation
  - Leverages generative power of GANs
  - Does not require stereo image pairs for training
  - Modular (plug any depth estimation method)

- Pipeline disadvantages:
  - Computationally expensive
  - Slow (3 step process)

Q&A