DEN: Disentangling and Exchanging Network for Depth Completion

You-Feng Wu*, Vu-Hoang Tran†, Ting-Wei Chang‡, Wei-Chen Chiu†, and Ching-Chun Huang‡
• **Intro**
  • Challenges of Depth Estimation
  • Our Setting
  • Overview

• **Previous Works**
  • Depth Representation Related
  • Disentangling Network Related

• **Method**
  • Depth Representation
  • Network Architecture
  • Criterion Design

• **Experiment**
Challenges of Depth Estimation

Monocular Depth Estimation

Challenges
- Spatial Scale Offset
- RGB image texture influence
- Mixed Depth Pixel
Challenges of Depth Estimation

Monocular Depth Estimation

Input

Depth Estimation Model

Alhashim et al.[9]

Ground truth

Challenges

- Spatial Scale Offset
- RGB image texture influence
- Mixed Depth Pixel

Model predicts incorrectly due to scale difference compare to ground truth image.

Challenges of Depth Estimation

Monocular Depth Estimation

Challenges
- Spatial Scale Offset
- RGB image texture influence
- Mixed Depth Pixel

Model is affected by the painting on the wall and predict undesired result.

Input

Alhashim et al.[9]  Ground truth

Challenges of Depth Estimation

Monocular Depth Estimation

Challenges
- Spatial Scale Offset
- RGB image texture influence
- Mixed Depth Pixel

Input

Zhang et al.[13]

Ground truth

Undesired prediction between foreground and background

Challenges of Depth Estimation

Monocular Depth Estimation

Challenges
- Spatial Scale Offset
- RGB image texture influence
- Mixed Depth Pixel

We can observe this problem clearer in 3D projection

Our Setting

**Motivation**
We aim to design an algorithm that can leverage on knowledge of incomplete depth map generated by commercial depth camera for
- More **Accurate** results compare to monocular depth estimation
- More **Practical** than design or purchase a higher-level depth camera for better quality depth map

**Depth Completion**

**Fail Cases of Commercial Depth Camera**
- Significant depth difference between foreground and background (A) [24]
- Shiny, bright, transparent (B), and distant surfaces (C) [25]

[25] Reconstructing Scenes with Mirror and Glass Surfaces, TOG, 2018
Our Setting

Depth Completion

Dataset we used

- ScanNet [7]
  - Used on several 3D scene understanding task
  - Provide 1. color image, 2. incomplete depth image, and annotated with 3. ground truth depth and 4. surface normal, etc.
  - Train: 59743 pairs of data from 1000 scenes
  - Test: another 500 pairs from other scenes
Overview

Model Architecture
- Disentangled Representation Learning
- Domain Adaptation
- Feature exchange across domains

Depth Representation
- General Depth Representation
Overview

Model Architecture
- Disentangled Representation Learning
- Domain Adaptation
- Feature exchange across domains

Depth Representation
- General Depth Representation

Challenges
- Spatial Scale Offset
- RGB image texture influence
- Mixed Depth Pixel
Overview

Model Architecture
- Disentangled Representation Learning
- Domain Adaptation
- Feature exchange across domains

Depth Representation
- General Depth Coefficient
Overview

Model Architecture
• Disentangled Representation Learning
• Domain Adaptation
• Feature exchange across domains

Depth Representation
• General Depth Coefficient

Challenges
- Spatial Scale Offset
- RGB image texture influence
- Mixed Depth Pixel
Previous Works
Depth Representation Related

Mixed Depth Pixel
- Many methods model depth estimation as a **regression problem** [12,18,20,35,44], which the model will prefer to generate mixed depth pixel for optimization.

[18] Depth Completion with Deep Geometry and Context Guidance, ICRA, 2019
[35] Dfusenet: Deep fusion of rgb and sparse depth information for image guided dense depth completion, arXiv, 2019
[44] Parse geometry from a line: Monocular depth estimation with partial laser observation, ICRA, 2017
Depth Representation Related

**DORN [4]**
- Depth Estimation: Regression Problem → Bin Classification Problem
- Loss: MSE/MAE → Cross Entropy

- Quantization Error
- Trade off between memory and precision

Depth Representation Related

DORN [4]
- Depth Estimation: Regression Problem → Bin Classification Problem
- Loss: MSE/MAE → Cross Entropy
- Quantization Error
- Trade off between memory and precision

Imran et al. [22]
- Proposed **Depth Coefficient** representation
- \( d = \alpha \times D_{k-1} + 0.5 \times D_k + \beta \times D_{k+1} \)
Disentangling Network Related

**DRIT [41]**
- Disentangling feature into **content** and **attribute** domain
- Aligning only **content** domain
- Exchanging **attribute** domain for style transferring

**Comparison**

![Diagram](image.png)

- **CycleGAN [38]**
- **DRIT [41]**
- **Our**

[38] Unpaired image-to-image translation using cycle-consistent adversarial networks, ICCV 2017
[41] Diverse image-to-image translation via disentangled representations, ECCV 2018
Method
Depth Representation

General Depth Coefficient

- Re-formula depth coefficients [22] based on spacing-increasing discretization (SID) [4]
  - Uniform Discretization (UD) [22]
    \[ UD: t_i = d_L + (d_U - d_L) \times \frac{i}{K} \]
  - Spacing Increasing Discretization (SID) [4]
    \[ SID: t_i = e^{\log(d_L) + \frac{\log(d_U - d_L) \times i}{\kappa}} \]

Depth Value → Depth Coefficient

\[ c_i = \{0, \ldots, 0, \alpha, 0.5, \beta, 0, \ldots, 0\} \]

where \( \alpha = \frac{d_i - 0.5(D_K + D_{K+1})}{D_{K-1} - D_{K+1}} \)
\( \beta = 0.5 - \alpha \)

[22] Depth Coefficients for Depth Completion, arXiv, 2019
**Depth Representation**

**General Depth Coefficient**
- Re-formula depth coefficients [22] based on spacing-increasing discretization (SID) [4]
  - Uniform Discretization (UD) [22]
    \[ UD: t_i = d_L + (d_U - d_L) \cdot i/K \]
  - Spacing Increasing Discretization (SID) [4]
    \[ SID: t_i = e^{\log(d_L) + \frac{\log(d_U - d_L) \cdot i}{K}} \]

**Depth Coefficient → Depth Value**
Network Architecture
Network Architecture

- Disentangled Representation Learning
- Domain Adaptation
- Feature exchange across domains
Network Architecture

- Disentangled Representation Learning
- Domain Adaptation
- Feature exchange across domains

Fail if force the model to align content domain directly

\[ L_{adv}^{content}(D_M) = \mathbb{E}_{C_S} \left[ \frac{1}{2} \log(D_M(C_S)) + \frac{1}{2} \log(1 - D_M(C_S)) \right] + \mathbb{E}_{C_T} \left[ \frac{1}{2} \log(D_M(C_T)) + \frac{1}{2} \log(1 - D_M(C_T)) \right] \]
Network Architecture

- Disentangled Representation Learning
- **Domain Adaptation**
- Feature exchange across domains

Learn the structural common content from both color and depth domain

The specific information for reconstruction
Network Architecture

- Disentangled Representation Learning
- Domain Adaptation
- Feature exchange across domains
Criterion Design

- $L_{\text{rec}}$
- $G_S(C_S, C_s^S, S_S)$
- $L_{\text{dom}}$
- $L_{\text{adv}}$
- $F_M(C_T)$
- $F_M(C_S)$
- $L_{\text{rec}}$
- $L_{\text{cycle}}$
- $L_{\text{mask}}$
- $L_{\text{latent}}$
- $X_S$
- $E_S^S$
- $S_S$
- $G_S$
- $C_S^S$
- $C_S^c$
- $E_S^c$
- $C_T^c$
- $C_T^S$
- $G_T$
- $X_T$
- $L_{\text{adv}}$
- $L_{\text{content}}$
- $E_T^c$
- $E_T^S$
- $S_T$
- $G_T(C_T, C_T^S, S_T)$
Criterion Design

What kind of loss do we need?

- $L_{rec}$: $l_i^{rec}(m_i, d_i, \hat{d}_i) = -m_i(d_i - \hat{d}_i)^2$
- $L_{ce}$: $l_i^{ce}(m_i, c_{ij}, \hat{c}_{ij}) = -m_i \sum_{j=1}^{K} c_{ij} \log \hat{c}_{ij}$

Reconstruction Loss
Criterion Design

What kind of loss do we need?

- $L_{rec}$: $l^{rec}_i(m_i, d_i, \hat{d}_i) = -m_i(d_i - \hat{d}_i)^2$
- $L_{ce}$: $l^{ce}_i(m_i, c_{ij}, \hat{c}_{ij}) = -m_i \sum_{j=1}^{k} c_{ij} \log \hat{c}_{ij}$

Reconstruction Loss
Surface Normal Loss
What kind of loss do we need?

- $L_{\text{rec}}$: $l_i^{\text{rec}}(m_i, d_i, \hat{d}_i) = -m_i(d_i - \hat{d}_i)^2$
- $L_{\text{ce}}$: $l_i^{\text{ce}}(m_i, c_{ij}, \hat{c}_{ij}) = -m_i \sum_{j=1}^{K} c_{ij} \log \hat{c}_{ij}$

Reconstruction Loss

Surface Normal Loss

Adversarial Loss

$$L_{\text{adv}}(D_M) = \mathbb{E}_{\mathcal{C}_S} \left[ \frac{1}{2} \log(D_M(C_S)) + \frac{1}{2} \log(1 - D_M(C_S)) \right] + \mathbb{E}_{\mathcal{C}_T} \left[ \frac{1}{2} \log(D_M(C_T)) + \frac{1}{2} \log(1 - D_M(C_T)) \right]$$
Criterion Design

What kind of loss do we need?

- $L_{\text{rec}}$: $l^\text{rec}_i(m_i, d_i, \hat{d}_i) = -m_i(d_i - \hat{d}_i)^2$
- $L_{\text{ce}}$: $l^\text{ce}_i(m_i, c_{ij}, \hat{c}_{ij}) = -m_i\sum_{j=1}^{K} c_{ij}\log \hat{c}_{ij}$

Reconstruction Loss
Surface Normal Loss
Adversarial Loss
Cycle Loss (L2-loss)
Experiment
Results

Comparison

Ours

Zhang et al. [13]

Much Less Mixed Depth Pixel

Example depth completion results on ScanNet test set.

Results

Comparison

Ours

Zhang et al.[13]

Much Less Spatial Scale Offset


Example depth completion results on ScanNet test set.
## Results

### Comparison

<table>
<thead>
<tr>
<th>Instance</th>
<th>Input</th>
<th>Ground truth</th>
<th>Ours</th>
<th>Zhang et al.[13]</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
</tr>
</tbody>
</table>

Point cloud visualization of our method and other comparisons. We convert the completed depth into point cloud.

## Results

### Comparison

<table>
<thead>
<tr>
<th>Obs.</th>
<th>Method</th>
<th>REL ↓</th>
<th>RMSE↓</th>
<th>1.25↑</th>
<th>1.25² ↑</th>
<th>1.25³ ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Zhang et al.[13]</td>
<td>0.0100</td>
<td>0.0155</td>
<td>0.9213</td>
<td>0.9588</td>
<td>0.9764</td>
</tr>
<tr>
<td></td>
<td>Ours(GDC)</td>
<td>0.0085</td>
<td>0.0132</td>
<td>0.9247</td>
<td>0.9621</td>
<td>0.9794</td>
</tr>
<tr>
<td>Y</td>
<td>Zhang et al.[13]</td>
<td>0.0076</td>
<td>0.0117</td>
<td>0.9588</td>
<td>0.9757</td>
<td>0.9856</td>
</tr>
<tr>
<td></td>
<td>Ours(GDC)</td>
<td>0.0063</td>
<td>0.0096</td>
<td>0.9617</td>
<td>0.9786</td>
<td>0.9877</td>
</tr>
<tr>
<td>N</td>
<td>Zhang et al.[13]</td>
<td>0.0408</td>
<td>0.0637</td>
<td>0.8113</td>
<td>0.9092</td>
<td>0.9492</td>
</tr>
<tr>
<td></td>
<td>Ours(GDC)</td>
<td>0.0386</td>
<td>0.0590</td>
<td>0.8160</td>
<td>0.9134</td>
<td>0.9551</td>
</tr>
</tbody>
</table>

Comparison against state-of-the-art algorithm on ScanNet dataset.
(B: GDT>0, Y: GDT>0 & RAW>0, N: GDT>0 & RAW=0)

*Best result show in blue.*
Thank You For Listening