Sensing, Understanding, and Synthesizing Humans in an Open World

Ziwei Liu

Nanyang Technological University
Human-Centric AI
Robust Sensing  
CelebA-Spoof

Synthesizing across Modalities  
Sep-Stereo

Understanding beyond Recognition  
Placepedia

Open World Learning  
Open Compound Domain Adaptation
Robust Sensing
CelebA-Spoof

Synthesizing across Modalities
Sep-Stereo

Understanding beyond Recognition
Placepedia

Open World Learning
Open Compound Domain Adaptation
CelebA-Spoof: Large-Scale Face Anti-Spoofing Dataset With Rich Annotations

Yuanhan Zhang*1,2  Zhenfei Yin*2  Yidong Li1  Guojun Yin2  Junjie Yan2  Jing Shao2  Ziwei Liu3

1Beijing Jiaotong University  2SenseTime Research  3The Chinese University of Hong Kong

*Equal contribution
Introduction
### CelebA-Spoof Dataset

1. Lack of Diversity  
2. Lack of Annotations

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Year</th>
<th>#Subjects</th>
<th>#Data(V/I)</th>
<th>#Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replay-Attack</td>
<td>2012</td>
<td>50</td>
<td>1,200 (V)</td>
<td></td>
</tr>
<tr>
<td>CASIA-MFSD</td>
<td>2012</td>
<td>50</td>
<td>600 (V)</td>
<td></td>
</tr>
<tr>
<td>3DMAD</td>
<td>2014</td>
<td>14</td>
<td>255 (V)</td>
<td></td>
</tr>
<tr>
<td>MSU-MFSD</td>
<td>2015</td>
<td>35</td>
<td>440 (V)</td>
<td></td>
</tr>
<tr>
<td>Msspoof</td>
<td>2015</td>
<td>21</td>
<td>4,704 (V)</td>
<td></td>
</tr>
<tr>
<td>HKBU-MARs V2</td>
<td>2016</td>
<td>12</td>
<td>1,008 (V)</td>
<td></td>
</tr>
<tr>
<td>MSU-USSA</td>
<td>2016</td>
<td>1,140</td>
<td>10,260 (I)</td>
<td></td>
</tr>
<tr>
<td>Oulu-NPU</td>
<td>2017</td>
<td>55</td>
<td>5,940 (V)</td>
<td></td>
</tr>
<tr>
<td>SiW</td>
<td>2018</td>
<td>165</td>
<td>4,620 (V)</td>
<td></td>
</tr>
<tr>
<td>CASIA-SURF</td>
<td>2018</td>
<td>1,000</td>
<td>21,000 (V)</td>
<td></td>
</tr>
<tr>
<td>CSMAD</td>
<td>2018</td>
<td>14</td>
<td>260 (V), 17 (I)</td>
<td></td>
</tr>
<tr>
<td>HKBU-MARs V1 +</td>
<td>2018</td>
<td>12</td>
<td>180 (V)</td>
<td></td>
</tr>
<tr>
<td>SiW-M</td>
<td>2019</td>
<td>493</td>
<td>1,628 (V)</td>
<td></td>
</tr>
<tr>
<td>CelebA-Spoof</td>
<td>2020</td>
<td>10,177</td>
<td>625,537 (I)</td>
<td>43</td>
</tr>
</tbody>
</table>
# CelebA-Spoof Dataset

## Presentation Attack Instrument

<table>
<thead>
<tr>
<th>Print</th>
<th>Paper Cut</th>
<th>Replay</th>
<th>3D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photo</td>
<td>Face</td>
<td>PC</td>
<td>Mask</td>
</tr>
<tr>
<td>Poster</td>
<td>Upper Body</td>
<td>Pad</td>
<td></td>
</tr>
<tr>
<td>A4</td>
<td>Region</td>
<td>Phone</td>
<td></td>
</tr>
</tbody>
</table>

## Illumination Condition and Environment

<table>
<thead>
<tr>
<th>Normal</th>
<th>Strong</th>
<th>Back</th>
<th>Dark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indoor</td>
<td>Indoor</td>
<td>Indoor</td>
<td>Indoor</td>
</tr>
<tr>
<td>Outdoor</td>
<td>Outdoor</td>
<td>Outdoor</td>
<td>Outdoor</td>
</tr>
</tbody>
</table>
Auxiliary Information Embedding Network (AENet)

Input image (Spoof) → CNN → Semantic → Classification → Geometric

Input image (Live) → Big Nose → Smile → Live
Observation
Depth Maps are More Versatile

- Smaller APCERg is better

- Reflection map as auxiliary supervision
- Depth map as auxiliary supervision

Depth Maps are More Versatile
Semantic Attribute Matters

Smaller $\text{APCER}_{S^s}$ and $\text{BPCER}_{S^f}$ are better

- $\text{AENet}_{C,S}$
- $\text{AENet}_{C,SW/oS^s}$
- $\text{AENet}_{C,SW/oS^f}$

Semantic Attribute Matters
Benchmark
Benchmarks

Intra-Dataset

Cross-Domain

Cross-Dataset

Training Data

Test Data
## Benchmarks

### Intra-dataset Benchmark

<table>
<thead>
<tr>
<th>Model</th>
<th>Backbone</th>
<th>Param. (MB)</th>
<th>Recall (%)(\uparrow)</th>
<th>AUC(\uparrow)</th>
<th>EER (%)(\downarrow)</th>
<th>APCER (%)(\downarrow)</th>
<th>BPCER (%)(\downarrow)</th>
<th>ACER (%)(\downarrow)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auxiliary* [23]</td>
<td>-</td>
<td>22.1</td>
<td>97.3</td>
<td>95.2</td>
<td>83.2</td>
<td>0.9972</td>
<td>1.2</td>
<td>5.71</td>
</tr>
<tr>
<td>BASN [16]</td>
<td>VGG16</td>
<td>569.7</td>
<td>98.9</td>
<td>97.8</td>
<td>90.9</td>
<td>0.9991</td>
<td>1.1</td>
<td>4.0</td>
</tr>
<tr>
<td>AENet(_{C,S,G})</td>
<td>ResNet-18</td>
<td>42.7</td>
<td><strong>98.9</strong></td>
<td>97.3</td>
<td>87.3</td>
<td>0.9989</td>
<td>0.9</td>
<td><strong>2.29</strong></td>
</tr>
</tbody>
</table>

### Cross-Domain Benchmark

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Model</th>
<th>Recall (%)(\uparrow)</th>
<th>AUC(\uparrow)</th>
<th>EER (%)(\downarrow)</th>
<th>APCER (%)(\downarrow)</th>
<th>BPCER (%)(\downarrow)</th>
<th>ACER (%)(\downarrow)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>93.7</td>
<td>86.9</td>
<td>69.6</td>
<td>0.996</td>
<td>2.5</td>
<td>5.7</td>
</tr>
<tr>
<td></td>
<td>AENet(_{C,G})</td>
<td>93.3</td>
<td>88.6</td>
<td><strong>74.0</strong></td>
<td>0.994</td>
<td>2.5</td>
<td>5.28</td>
</tr>
<tr>
<td></td>
<td>AENet(_{C,S})</td>
<td>93.4</td>
<td>89.3</td>
<td>71.2</td>
<td>0.996</td>
<td>2.4</td>
<td>5.63</td>
</tr>
<tr>
<td>AENet(_{C,S,G})</td>
<td>95.0</td>
<td>91.4</td>
<td>73.0</td>
<td>0.995</td>
<td>2.1</td>
<td>4.09</td>
<td>2.09</td>
</tr>
</tbody>
</table>

### Cross-Dataset Benchmark

<table>
<thead>
<tr>
<th>Model</th>
<th>Training</th>
<th>Testing</th>
<th>HTER (%)(\downarrow)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAS-TD-SF [34]</td>
<td>SiW</td>
<td>CASIA-MFSD</td>
<td>39.4</td>
</tr>
<tr>
<td>FAS-TD-SF [34]</td>
<td>CASIA-SURF</td>
<td>CASIA-MFSD</td>
<td>37.3</td>
</tr>
<tr>
<td>AENet(_{C,S,G})</td>
<td>SiW</td>
<td>CASIA-MFSD</td>
<td><strong>27.6</strong></td>
</tr>
<tr>
<td>Baseline</td>
<td>CelebA-Spoof</td>
<td>CASIA-MFSD</td>
<td>14.3</td>
</tr>
<tr>
<td>AENet(_{C,G})</td>
<td>CelebA-Spoof</td>
<td>CASIA-MFSD</td>
<td>14.1</td>
</tr>
<tr>
<td>AENet(_{C,S})</td>
<td>CelebA-Spoof</td>
<td>CASIA-MFSD</td>
<td>12.1</td>
</tr>
<tr>
<td>AENet(_{C,S,G})</td>
<td>CelebA-Spoof</td>
<td>CASIA-MFSD</td>
<td><strong>11.9</strong></td>
</tr>
</tbody>
</table>
CelebA-Spoof
Large-Scale
Collection Dimensions
<table>
<thead>
<tr>
<th>Angle</th>
<th>Input Sensor</th>
<th>Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Rich Annotations
<table>
<thead>
<tr>
<th>Illumination Condition</th>
<th>Spoof Type</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Photo Face Mask</td>
<td>Indoor</td>
</tr>
<tr>
<td>Strong</td>
<td>Poster Face Mask</td>
<td></td>
</tr>
<tr>
<td>Dark</td>
<td>A4 Upper Body Mask</td>
<td></td>
</tr>
<tr>
<td>Dark</td>
<td>Region Mask PC</td>
<td>Outdoor</td>
</tr>
<tr>
<td>Dark</td>
<td>Face Mask Tablet</td>
<td></td>
</tr>
<tr>
<td>Back</td>
<td>Phone 3D Mask Phone</td>
<td></td>
</tr>
</tbody>
</table>
Face Mask
Upper Body Mask
Region Mask
Github Page: https://github.com/Davidzhangyuanhan/CelebA-Spoof
Robust Sensing
CelebA-Spoof

Synthesizing across Modalities
Sep-Stereo

Understanding beyond Recognition
Placepedia

Open World Learning
Open Compound Domain Adaptation
Placepedia: Comprehensive Place Understanding with Multi-Faceted Annotations

Huaiyi Huang, Yuqi Zhang, Qingqiu Huang, Zhengkui Guo, Ziwei Liu, Dahua Lin

The Chinese University of Hong Kong
Dataset Overview

• Place is an important element in visual understanding.
  • Functionality
  • Cultural style
  • Economic type

• Comprehensive place understanding
  • Far beyond categorizing a place with an image
  • Requires information of multiple aspects
Dataset Overview

Musée du Louvre
Street-address: Place du Carrousel
Function: See
Category: Museums
Description: “Its most famous exhibit, …, is Leonardo da Vinci’s painting of the Mona Lisa, …”
Coordinates: 48.86°N 2.34°W
Opening Hours: W-M 09:00-18:00, closed public holidays ...
Price: €12-16; under 18, free ...

1st arrondissement

Île-de-France

7th arrondissement

La Tour Eiffel
Nickname: The Eiffel Tower
Function: See
Category: Landmarks
Description: “A symbol of Paris and one of the most famous landmarks in the world. …”
Coordinates: 48.86°N 2.29°W
Opening Hours: 09:30-23:45; Jun 21-Sep 02: 09:00-00:45; Jul 14, Jul 15 off
Price: 25€ (12€ for age 12-24) …

Paris
Description: “… has the reputation of being the most beautiful and romantic of all cities, …”
Population: 2,140,526
Area: 40.7 sq mi
Time zone: UTC+1
Elevation: 114 ft
Coordinates: 48.86°N 2.35°E
Get Very Rich Data of Place

• Wikivoyage!
• 360K places of 25K cities
• Lots of meta data
Get Very Rich Data of Place

• Some are not place. Refine the place list!

  • It has the attribute GPS coordinates or address
  • It is identified as a location by Google Entity Recognition [1] or Stanford Entity Recognition [2]

    • Grant Plaza Hotel©, 485 Grant Ave 5th Pine St., +1-415-434-3383, fsc: +1-412
      2 night stay. Rates are reasonable. Make sure you ask for one of the outside rooms
      Is a place.

    • Chinese New Year Festivities©, Jan or Feb. Celebrated for over 5,000 years,
      firecrackers, "lucky-money" envelopes, colorful banners, over 100 ornately thi
      Is not a place.

• Remains 320K places

Get Very Rich Data of Place

- Google Image open source
- 240K places
- 35M images
Challenges

| Pan Pacific Vancouver Canada Splurge Hotels Sleep | Pat's King of Steaks Philadelphia Region United States of America Restaurants Eat | Cafe Tortoni Buenos Aires Argentina Cafe Drink | Red Pyramid Cairo Egypt Landmarks See | Galaxy Macau Macau Gambling Do | Rangitoto Island Auckland New Zealand Islands See & Do | Sydney Fish Market Sydney Australia Markets Buy | Oslo Metropolitan University Oslo Norway Universities Learn | Dublin Airport Dublin Ireland Airports Get in | Airport Rail Link Bangkok Thailand Public Transportations Get around |

1) Daytime to nighttime  
2) Different angles  
3) Inside and Outside
Comprehensive Place Understanding

• Benchmarks

  • Datasets:

    • Places-Coarse: 26K places
    • Places-Fine: 1K places

  • Tasks:

    • Place Retrieval (determine if two images belong the same place)
    • Place Categorization (classify places into categories like museums, parks, churches, and temples)
    • Function Categorization (classify places by their functionality such as eat, sleep, see, buy, and so on)
    • City/Country Recognition (classify places into their cities or countries)
A unified Framework to Predict All Tasks

- Duplicate the last convolution/pooling/fc layers of ResNet50 to five branches
City Embedding

• For vision:
  • Using place images
  • Extract the feature from the city recognition model

• For text:
  • Using city descriptions
  • Embed the content of texts into numeric space based on Bert pre-defined model

• For both vision & text: concatenate two vectors above
City Embedding Visualization
City Embedding

- City description
- Calculate the weights
  - Economic
  - Cultural
  - Political
  as in [1]
- Pearson correlation
- Compare with neuron

Conclusion

• A large-scale place dataset
  • comprehensively annotated with multiple aspects

• Explore place understanding
  • Build several benchmarks and study a unified model to recognize places
  • Remains lots of challenges

• Learn city embedding representations
  • Learning from both visual & textual domains can better characterize a city
  • Economic/cultural/political elements could be expressed in different types of images
Placepedia: Comprehensive Place Understanding with Multi-Faceted Annotations

Project page: https://hahehi.github.io/placepedia.html

Code and models: https://github.com/hahehi/placepedia
Robust Sensing
CelebA-Spoof

Understanding beyond Recognition
Placepedia

Synthesizing across Modalities
Sep-Stereo

Open World Learning
Open Compound Domain Adaptation
Sep-Stereo: Visually Guided Stereophonic Audio Generation by Associating Source Separation

Hang Zhou*, Xudong Xu*, Dahua Lin, Xiaogang Wang, Ziwei Liu

CUHK – SenseTime Joint Lab, The Chinese University of Hong Kong
Prior Works: Fully Supervised Training

Spatial AudioGen (Morgado et al, NeurIPS 2018)

Spatial AudioGen exploits the stereophonic data on YouTube.

2.5d Visual Sound (Gao et al, CVPR 2019)

Mono2Binaural relies on self-collected dataset, FAIR-Play, to train the network.
Limitation: Stereophonic Data Collection

Data collection equipment in:
2.5d Visual Sound

(a) Dummy Head Recording

(b) Binaural Microphone
Mono Audios in Source Separation

The Sound of Pixels
(ECCV 2018)

Co-Separating Sounds of Visual Objects
(ICCV 2019)

• Massively availed mono audios have been successfully used in the field of source separation.
Key Insight: Regard the problem of separating two audios as an extreme case of creating binaural audio.
Key Insight: Regard the problem of separating two audios as an extreme case of creating binaural audio.

- Mono audios can be used to facilitate the generation of stereo audio.
A Unified Framework: Sep-Stereo
Stereophonic Learning
Stereophonic Learning

**Training Data Settings:** Same as 2.5D Visual Sound.

**Base Network:** Same as 2.5D Visual Sound.

**Visual Feature:** Kept as 14x7 feature map.
Separative Learning

(a) Stereophonic Learning

(b) Separative Learning

$V_b$

$V_a$

$V_s$

$a_{mix}$

$a_{mono}$

$S_{mix}$

$S_{mono}$

$Net_v$

MaxPool

Res18

$F_v^0$

Transfer

Associative Pyramid Network

$S_b^p$

$S_a^p$

$S_D$

$S_r^p$

$S_l^p$
Separative Learning

**Training Data Settings:** Same as Sound of Pixels.

**Base Network:** Same as 2.5D Visual Sound.

**Visual Feature:** Max-pooled and rearranged.
Difference: Visual Feature Rearrangement

Stereophonic Learning

Separative Learning
Associative Pyramid Network
Associative Pyramid Network

- Associative Pyramid Network better associates the visual features and the audio features with a learned Associative-Conv operation.

\[
F_{ap}^i = \text{Conv2d}(F_{ap}^i) \\
F_{ap}^i = \text{Cat}([\text{DeConv}(F_{ap}^{i-1}), F_{ap}^i])
\]
Mono audio helps stereo audio generation

- Extensive experiments demonstrate that Sep-Stereo can achieve better performance on the task of stereo audio generation with the help of mono audio data.

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Data</th>
<th>FAIR-Play</th>
<th>YT-Music</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>STFT$_D$</td>
<td>ENV$_D$</td>
</tr>
<tr>
<td>Mono2Binaural</td>
<td>✓</td>
<td>0.959</td>
<td>0.141</td>
</tr>
<tr>
<td>Baseline (MUSIC)</td>
<td>✓</td>
<td>0.930</td>
<td>0.139</td>
</tr>
<tr>
<td>Associative-Conv</td>
<td>✓</td>
<td>0.893</td>
<td>0.137</td>
</tr>
<tr>
<td>APNet</td>
<td>✓</td>
<td>0.889</td>
<td>0.136</td>
</tr>
<tr>
<td>Sep-Stereo (Ours)</td>
<td>✓</td>
<td>0.879</td>
<td>0.135</td>
</tr>
</tbody>
</table>
Demo Results

For **stereo**, we compare our results with Mono2Binaural (model of 2.5D Visual Sound), we show the mono input, results from two models and the ground truth.

For **separation**, we compare with PixelPlayer (Sound of Pixels).

Results on MUSIC duets demonstrate the generalization of our method.
Better watch with HIGH QUALITY earphones or headphones
Future Work

• Separation:
  • Universal separation, tackling music and speech, even general sound with one model.
  • Adopt the ideas from state-of-the-art audio source separation for the pursuit of models with more capacity.
  • Exploring the task of visually guided audio generation and separation together.

• Stereo:
  • The problem of overfitting still remains unsolved.
  • How to incorporate the setting of the room into the generation of stereo.
Project page: https://hangz-nju-cuhk.github.io/projects/Sep-Stereo

Code and models: https://github.com/SheldonTsui/SepStereo_ECCV2020
Robust Sensing
CelebA-Spoof

Understanding beyond Recognition
Placepedia

Synthesizing across Modalities
Sep-Stereo

Open World Learning
Open Compound Domain Adaptation
Open Compound Domain Adaptation

Ziwei Liu* Zhongqi Miao* Xingang Pan Xiaohang Zhan Dahua Lin Stella X. Yu Boqing Gong

The Chinese University of Hong Kong UC Berkeley / ICSI Google Inc.
Perception for Autonomous Driving
Compound Heterogeneous Domains

- Cloudy
- Rainy
- Snowy

Simulation

Source

Target
Simulation

Source domain

Rainy

(a) Unsupervised Domain Adaptation

Rainy

(b) Multi-Target Domain Adaptation

Rainy

Snowy

Overcast

Unseen weather and more

Open Compound Domain Adaptation
Challenges:

1) **Compound Heterogeneous Domains**
   -> Traditional DA works on pairwise adaptation settings

1) **Open Unknown Domains**
   -> Traditional DA assumes prior access to domain data during training
Open Compound Domain Adaptation

A compound target domain

Unseen weather and more

Rainy

Snowy

Overcast

Cloudy

Challenges:
1) Compound Heterogeneous Domains
   -> Traditional DA works on pairwise adaptation settings

1) Open Unknown Domains
   -> Traditional DA assumes prior access to domain data during training

Architecture

memory

Learning

curriculum

domain disentanglement

adaptive knowledge transfer
Continuous Adaptation
Continuous Adaptation

Source
Simulation

Compound Targets
Open World Driving Conditions
Cloudy
Rainy

Open Targets
Overcast
domain memory

instance-wise curriculum

domain memory
Source
Simulation

Compound Targets
Open World Driving Conditions
Cloudy
Rainy

Open Targets
Overcast

Domain Disentanglement
instance-wise curriculum
Continuous Adaptation

Adaptive Knowledge Transfer
domain memory
Continuous Adaptation

Source

SVHN

Compound Targets

Open Compound Domain Digits Classification

MNIST-M

MNIST

USPS

Open Targets

SymNum

Domain Disentanglement

instance-wise curriculum

Adaptive Knowledge Transfer

Domain memory
Adversarial Domain Characteristics
Disentanglement

\[
\min_{E_{\text{domain}}} \quad - \sum_i z^i_{\text{random}} \log D(E_{\text{domain}}(x^i)) \\
\min_D \quad - \sum_i y^i \log D(E_{\text{domain}}(x^i))
\]
Source

Domain Disentanglement

Compound Targets
instance-wise curriculum

Open Targets
domain memory

Continuous Adaptation

Adaptive Knowledge Transfer
Curriculum according to Domain Characteristics

- Domain “SVHN” (source)
- Domain “MNIST-M” (compound)
- Domain “USPS” (compound)
- Domain “MNIST” (compound)

Sample %

Epoch 1 → Epoch K

Source

Domain Disentanglement

Compound Targets
instance-wise curriculum

Open Targets
domain memory

Continuous Adaptation

Adaptive Knowledge Transfer
Memory-Augmented Domain Indicator

\[ v_{\text{transfer}} = v_{\text{direct}} + \epsilon_{\text{domain}} \otimes v_{\text{enhance}} \]

Source

Domain Disentanglement

Compound Targets
instance-wise curriculum

Continuous Adaptation

Open Targets
domain memory

Adaptive Knowledge Transfer
C-Digits Benchmark
Absolute Performance Gain: ~5%

C-Faces Benchmark
Absolute Performance Gain: ~10%

C-Driving Benchmark
Absolute Performance Gain: ~2%

C-Mazes Benchmark
Absolute Performance Gain: ~30%
Robustness to the complexity of compound domains and open domains
Adaptation Results on C-Driving

(semantic segmentation)
Compound Target Domain (Rainy)

Source Only

Ours
Open Target Domain (Overcast)

Source Only

Ours
Adaptation Results on C-Mazes

(reinforcement learning)
Open Target Domain 1

MTL (fail)

SynPo (succeed)

Ours (succeed)
Open Target Domain 2

MTL (fail)

SynPo (fail)

Ours (succeed)
New Task
Open Compound Domain Adaptation (OCDA)

New Approach
Instance-wise Curriculum + Domain Memory

New Benchmarks
C-Digits, C-Faces, C-Driving, and C-Mazes
Code, models and benchmarks are available at

What’s Next

“Devils are in the Tails”  “Blessing of Dimension”  “Ghost in the Shell”
Thanks!

Science is what we understand well enough to explain to a computer. Art is everything else we do.

Homepage: https://liuziwei7.github.io/