# Enhancing the performance of image classification through features automatically learned from depth-maps

George Ciubotariu, Vlad-Ioan Tomescu, Gabriela Czibula

September 2021

# Contents

# Research Questions and Original Contributions

- **RQ1**: *How relevant are depth maps in the context of indoor-outdoor image classification?*
  - Unsupervised learning based analysis on DIODE dataset for indoor-outdoor classification
    - t-SNE clustering support for further supervised investigations
- **RQ2**: *To what extent does aggregating visual features into more granular sub-images increase the performance of classifiers?*
  - Supervised learning based classification for supporting the unsupervised approach
    - Multilayer Perceptron (MLP) classifier tested to confirm hypothesis
- **RQ3**: *How correlated are the results of the unsupervised based analysis and the performance of supervised models applied for indoor-outdoor image classification?*
  - Comparative analysis on image features aggregation

# Introduction in the Approached Tasks

- Indoor-Outdoor Classification
  - motivation
- Semantic Segmentation
- Depth Estimation

# Related Work

- A review on indoor-outdoor scene classification, feature extraction methods, classifiers and data sets is done by Tong et al. [TSYW06]
  - multiple remarkable methods
  - mentions good performances between 1998 and 2017
  - features such as color, texture, edge etc.
  - multiple data sets were mentioned
- Cvetkovic et al. [CNI14]
  - color and texture descriptors and a SVM classifier
  - results of 93.71% and 92.36% accuracy on two public data sets
- Tahir et al. [TMR15]
  - computes the GIST descriptor as a feature vector
  - 90.8% accuracy on a public data set
- Raja et al. [RRDR13]
  - uses HSV instead of RGB color encoding
  - extracts color, texture and entropy features
  - features extracted from 100 sub-images
  - lightweight KNN classifier

# Computer Vision (CV) and Deep Learning (DL)

Most recent work implement **Convolutional Neural Networks** (CNNs) in dense visual tasks such as *Semantic Segmentation* (SS) or *Depth Estimation* (DE).

- [LRSK19, RBK21] **Dense Prediction Transformers** (DPT)
    - model that leverages visual transformers instead of convolutions.
    - robust architecture to serve as a backbone in our experiments
    - tested for both SS and DE tasks, achieving great results, therefore offering us the possibility to create a comparative approach

# Vision Transformers for Dense Prediction (DPT)

| Model | Image resolution | # extracted features after encoder | # extracted features after decoder |
|---|---|---|---|
| **Depth Estimation** | 384×384 | 49152 | 12582912 |
| **Semantic Segmentation** | | | |

Table: DPT architectures details

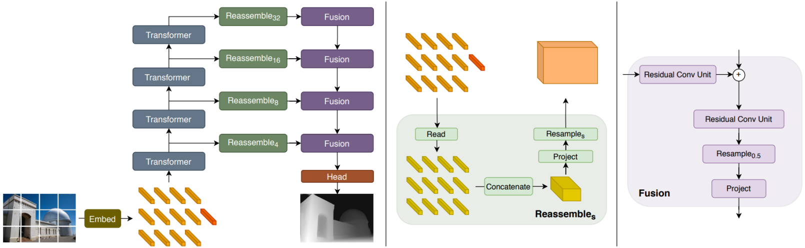


Figure: DPT architecture

# DIODE (Dense Indoor and Outdoor DEpth)

- Data has been collected with a **FARO Focus S350**
- It consists of 27858 $1024 \times 768$ **RGB-D** images
- Photos have been taken both at daytime and night, over several seasons (summer, fall, winter)

Apart from RGB-D images, DIODE dataset also provides us with normal maps that could further enhance the learning of depth and vice-versa

# DIODE (Dense Indoor and Outdoor DEpth)

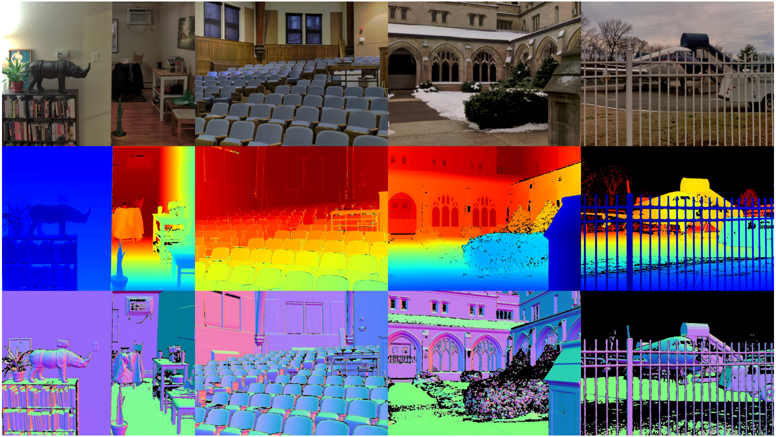

Figure: Sample images from DIODE dataset

# DIODE Structure

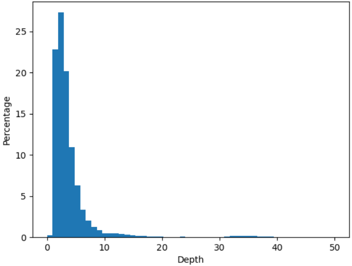


Figure: Histogram of depth values frequency (%) for indoor train set

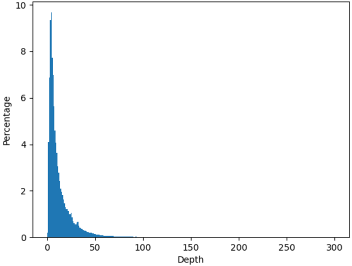


Figure: Histogram of depth values frequency (%) for outdoor train set

# Methodology

- Feature extraction
  - manually engineered features
  - automatically learned features
- Unsupervised learning-based analysis
- Supervised learning-based analysis
  - depth-augmented images

# Automatic Feature Extraction

1. **aggregating RGB from sub-images**
   - $3 \cdot k$ dimensional vector ($k = 1, 4, 16$)
   - average RGB values for each sub-image
2. **aggregating RGBD from sub-images**
   - $4 \cdot k$ dimensional vector ($k = 1, 4, 16$)
   - average RGBD values for each sub-image
3. **features from DPT encoder/decoder**
   - trained for SS
   - trained for DE

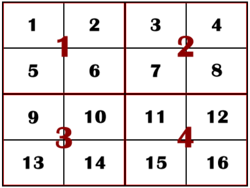


Figure: Structure of image splits

# Unsupervised Learning for Analysing the Data

- *3D t-SNE* unsupervised clustering
  - used for *non-linear* dimensionality reduction
  - able to uncover more useful patterns in data
  - uses *Student t-distribution* to better disperse the clusters
- *data normalization* with the **inverse hyperbolic sine (asinh)**
  - increased sensitivity to particularly small and large values
- parameters used
  - **perplexity** of 20
  - **learning rate** of 3.0
    - for a slower converging but finer learning curve
  - 1000 **iterations**

| **Measure** | **RGBD features (4 splits)** | **DPT DE learned features** | **DPT SS learned features** | **DPT SS depth augmented features** |
|---|---|---|---|---|
| *Prec* | 0.769 | 0.729 | **0.945** | **0.957** |

Table: *Prec* values for the t-SNE transformations depicted in Figures 6 – 9.

# Features extracted aggregating RGB and RGBD values

- 4 splits

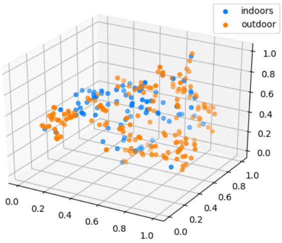


Figure: t-SNE for RGB with 4 splits

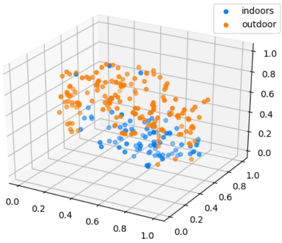


Figure: t-SNE for RGB-D with 4 splits

# Features Extracted from DL models

- DPT trained for Semantic Segmentation

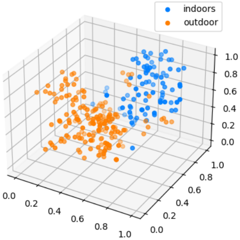


Figure: t-SNE of DPT encoder extracted features for SS

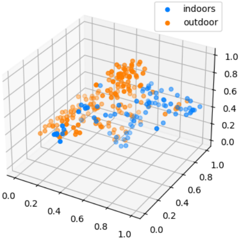


Figure: t-SNE of DTP encoder extracted features for DE

# Supervised Learning Results

| Features | # Splits ($n$) | Accuracy | AUC | Specificity | Sensitivity |
|---|---|---|---|---|---|
| RGB | 0 | 0.692±0.077 | 0.525±0.056 | 0.980±0.028 | 0.070±0.121 |
| | 1 | 0.688±0.064 | 0.517±0.022 | **0.989**±0.014 | 0.046±0.049 |
| | 2 | 0.669±0.049 | 0.545±0.048 | 0.912±0.068 | 0.163±0.136 |
| RGBD | 0 | **0.880**±0.039 | 0.858±0.041 | 0.898±0.058 | 0.817±0.081 |
| | 1 | 0.876±0.043 | **0.862**±0.044 | 0.894±0.046 | 0.829±0.063 |
| | 2 | 0.838±0.044 | 0.826±0.053 | 0.848±0.060 | 0.804±0.099 |
| DPT-DE | 0 | 0.823±0.131 | 0.831±0.076 | 0.812±0.185 | 0.850±0.069 |
| **DPT-SS** | 0 | **0.950**±0.027 | **0.942**±0.029 | 0.969±0.034 | **0.915**±0.053 |
| **DPT-SS+D** | 0 | **0.961**±0.015 | **0.956**±0.021 | **0.970**±0.019 | **0.941**±0.041 |

Table: The results of supervised learning indoor-outdoor classification on DIODE dataset. Confidence intervals of 95% were used in the analysis. Only the features extracted by the DPT encoder are used in the experiments.

# Comparative Results

Benefits of our method:

- lightweight
  - uses less features and parameters compared to other models
  - low memory and computational cost compared to other deep learning methods
  - significant increase in performance when adding depth cues
- capable of being optimised using multi-threading
- displays potential of depth cues use for multiple visual tasks

According to the study performed by Tong et al., our approach which uses features extracted using DPT-SS+D (96.1% accuracy) establishes a new State-of-the-art in indoor-outdoor classification. The best performance presented in [TSYW06] is 93.8% accuracy.

# Ongoing Experiments and Future Enhancements

- Identifying features that can be used in both SS and DE
- Identifying other problems that can be solved with adapted DL models
- Architecture Transfer from SS towards DE
- Multitask and Collaborative Learning

Thank you!

Questions?

# Bibliography I

Stevica Cvetkovic, Sasa Nikolic, and Slobodan Ilic.
Effective combining of color and texture descriptors for indoor-outdoor image classification.
*Facta universitatis - series: Electronics and Energetics*, 27:399–410, 01 2014.

Katrin Lasinger, René Ranftl, Konrad Schindler, and Vladlen Koltun.
Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer.
*CoRR*, abs/1907.01341, 2019.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun.
Vision transformers for dense prediction.
*CoRR*, abs/2103.13413, 2021.

# Bibliography II


R. Raja, S. Md. Mansoor Roomi, D. Dharmalakshmi, and S. Rohini.
Classification of indoor/outdoor scene.
In *2013 IEEE International Conference on Computational Intelligence and Computing Research*, pages 1–4, 2013.

Waleed Tahir, Aamir Majeed, and T. Rehman.
Indoor/outdoor image classification using gist image features and neural network classifiers.
*12th International Conference on High-capacity Optical Networks and Emerging Technologies*, pages 1–5, 2015.

Zhehang Tong, Dianxi Shi, Bingzheng Yan, and Jing Wei.
A review of indoor-outdoor scene classification.
In *Proceedings of the 2017 2nd International Conference on Control, Automation and Artificial Intelligence (CAAI 2017)*, pages 469–474. Atlantis Press, 2017/06.