19th CVC Workshop on Computer Vision Trends and Challenges

Proceedings of the nineteenth CVC Workshop CVCR&D2024

Centre de Visió per Computador
Bellaterra, Catalonia (Spain)
July 19, 2024

Copyright © 2024 by the authors in the table of contents. All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the authors.

General chairs

Mohamed Ali Souibgui
Matías Bilkis

Organization comitee

Núria Martínez
Marc Compte
Aurora García

Program comitee

Jordi Casas
Joaquín Cerdà
Javier Vazquez
Juan J. Villanueva
Fernando Vilariño
Maria Vanrell
Ernest Valveny
Joan Serrat
Francisco Sanchez
Gemma Sánchez
Francesc X. Roca
Oriol Ramos
Daniel Ponsa
Joost Van de Weijer
Jose Luis Gomez
Gabriel Villalonga
Bartlomiej Twardowski
Carlos Alejandro
Xavier Otazu
Felipe Lumbreras
Antonio M. López
Josep Lladós
Dimos Karatzas
Aura Hernandez
Jordi Gonzàlez
Debora Gil
Alicia Fornés
Jorge Bernal
Roberto Benavente
Ramon Baldrich
Mikhail Mozerov
Bogdan Raducanu
Angel Sappa
ai Wang
Najlaa Maroof
Andrey Barsky
Adrian Galdran
Lluis Gómez
Carles Sánchez
Ricardo Jorge
Julio Cezar
Jacques Junior
Eloi Puertas
Simone Balocco
Oliver Díaz
Albert Clapés
Sergio Escalera
Yi Xiao
Arka Ujjal Dey
Sandesh Kamath
Lei Kang

Table of contents

Preface 7

Author Index 8

1. Image analysis and enhancement

From spatial super-resolution to bit-depth recovery
Xuanshuo Fu 9

Implicit Neural Representation Model for Camera Relocalization with Learning in Global Multi-Scenes Scenario
Shun Yao 12

SR-AerIA: Super-Resolution for Aerial Imagery Application
Mohamed Ramzy Ibrahim 14

Color in Visual-Language Models: CLIP deficiencies
Guillem Arias 16

Anomaly Detection for Efficient Lesion Screening in Gastrointestinal Endoscopy
Yael Tudela Barroso 18

Semantic Segmentation in Multispectral Imaging
Leo Ramos 20

Rating symmetry in digital images: a biologically-inspired approach
Mehdi Fatan Serj 22

Leveraging Color Names for Image Enhancement
David Serrano-Lozano 27

Multi-Light Intrinsic-Aware Neural Radiance Fields
Yixiong Yang 29

2. Document image understanding

Privacy-Aware Document-based VQA
Khanh Nguyen 31

Intelligent Reading Systems for Real-World Applications
Sergi Garcia Bordils 33

Multi-page document VQA with Recurrent Memory Transformer
Qi Dong 35

Multi-Modal Structured Reasoning for Historical Document Information Extraction
Adrià Molina Rodríguez 36

Multimodal Models for Comics Understanding
Emanuele Vivoli 41

Computer Vision Techniques for Historical Handwritten Music Recognition
Pau Torras 43

Information spotting in sequences and time series: Application to gestures and strokes
Asma Bensalah 45

3. Advanced driver-assistance systems (ADAS)

Divide, Ensemble and Conquer: The Last Mile on Unsupervised Domain Adaptation for On-Board Semantic Segmentation
Lin Tao 46

Minimizing Human Labeling in Training Deep Models for Pedestrian Intention Prediction
Muhammad Naveed Riaz 48

UDA4Inst: Unsupervised Domain Adaptation for Instance Segmentation
Yachan Guo 50

Guiding Attention in End-to-End Driving Models
Diego Porres 52

Developing Transparent Artificial Intelligence Mechanisms for Enhanced Decision-Making in Autonomous Vehicles
Alexandre Levy 53

4. Continual Learning

Continually Learning Self-Supervised Representations
Alex Gomez-Villa *55*

Exemplar-Free Continual Learning
Dipam Goswami *57*

5. Image generation

Generative models in Cross-Domain Image Processing
Patricia L. Suárez *59*

ColorPeel: Color Prompt Learning with Diffusion Models via Color and Shape Disentanglement
Muhammad Atif Butt *61*

Sketching Imagination: Explorations of Generative Models for Visual Creativity
Ayan Banerjee *63*

Synthesizing Aesthetic Diversity: Creating a Balanced Dataset ofVisual Aesthetics Using GANs
Arslan Javed *64*

SR-AerIA: Super-Resolution for Aerial Imagery Application
Héctor Laria Mantecón *66*

Adversarial Attacks on Latent Diffusion Models
Xide Xu *67*

Preface

This book contains the papers presented at the nineteenth CVC Workshop on Computer Vision Trends and Challenges (CVCR&D'2024). This edition, the workshop was held in CVC on July 19, 2024.

The CVC workshops provide an excellent opportunity for young researchers and project engineers to share new ideas and knowledge about the progress of their work, and also, to discuss challenges and future perspectives. In addition, the workshop is a welcome event for those who have recently joined the Institute.

The CVCR&D program is organized in a single-track single-day workshop. It comprises several sessions dedicated to specific topics. For each session, a doctor working on the topic introduces the general research lines. The PhD students expose their specific research. A poster session will be held for open questions. Session topics cover the current research lines and development projects of CVC: Medical Imaging, Color & Texture Analysis, Object Recognition, Image Sequence Evaluation, Advanced Driver Assistance Systems, Machine Vision, Document Analysis, Pattern Recognition, Human Pose Recovery and Behavior Analysis, and Applications. We want to thank all paper authors and Program Committee members. Their contribution shows that CVC has a dynamic, active, and promising scientific community.

We hope you all enjoy this nineteenth workshop and we are looking forward to meeting you next year at the twentieth CVCR&D.

Bellaterra, July 2024
Mohamed Ali Souibgui & Matías Bilkis
Workshop General Chairs

Author Index

A. Banerjee 63
A. Gomez-Villa 55
A. Levy 53
A. Javed 64
A. Molina Rodríguez 36
D. Goswami 57
D. Porres 52
D. Serrano Lozano 27
E. Vivoli 41
G. Arias Bedmar 16
G. Yachan 50
H. Laria 66
K. Nguyen 31
L. Ramos 20
M. Atif Butt 61
M. Fatan Serj 22
M. Naveed Riaz 48
M. Ramzy Ibrahim Mahmoud 14
P. Leonor Suarez Riofrio 59
P. Torras 43
Q. Dong 35
S. Garcia Bordils 33
S. Yao 12
T. Lian 46
X. Fu 09
X. Xu 67
Y. Tudela Barroso 18
Y. Yang 29

From spatial super-resolution to bit-depth recovery

Xuanshuo Fu
Advisor: Javier Vazquez-Corral
Computer Vision Center & UAB
E-mail: xuanshuo@cvc.uab.es

Keywords: color image enhancement, bit depth recovery, color image restoration, explainability

1 Summary of Previous and Current Work

My current research investigates the application of deep learning techniques for reconstructing high-bit-depth images from their compressed, low-bit-depth counterparts, a problem usually call bit-depth recovery. In general terms, bit-depth recovery is a post-processing technique aiming at recovering contrast information lost during the image capture process. There are two different ways of performing bit-depth recovery: direct estimation and residual estimation. Direct estimation methods attempt to directly predict the high-bit-depth image from the low-bit-depth input. Residual estimation methods, on the other hand, focus on predicting the difference (residual) between the high and low-bit-depth images.

In particular, the aim of my current work is to leverage information from spatial super-resolution algorithms with the aim of using it as a guidance to recover bit-depth information, replacing spatial resolution by bit-depth resolution. Our approach works as a residual estimation method, where we estimate a different network for each extra bit we want to recover.

The current version of our model can be seen in Figure 1. In short, given an input image we first process it using a super-resolution network (we are currently using ESRGAN [1]) to obtain x2 and x4 versions of the same image. We also apply an Inception network to the original image. We then fuse all these outputs and give them to the baseline method —a method already designed for bit-depth super-resolution (D4 [2])— to obtain our result.

Preliminary results of our approach—including an ablation study— for the 4 to 8 bits case are shown in Table 1. We present results both in terms of PSNR and SSIM. As we can see both of our base cases —including only the inception model or the super-resolution— already improve the results. Best results are when both parts are combined together.

2 Future Work and Challenges

Future Work: As a future work we plan to test our method with different super-resolution networks and different baseline approaches to prove its generality.
Challenges: A big challenge is that our most effective current color recovery algorithm involves recovering one bit at a time and training a specific model for each bit. This approach allows each model to concentrate on the features relevant to its specific bit, avoiding confusion from focusing on too many features. However, this sequential recovery process increases the training time proportionally with the model's complexity. Looking ways to train the model is a less cumbersome manner will also be studied.

3 References

@inproceedingslim2017enhanced, title=Enhanced deep residual networks for single image super-resolution, author=Lim, Bee and Son, Sanghyun and Kim, Heewon and Nah, Seungjun and Mu Lee, Kyoung, booktitle=Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages=136–144, year=2017

Publications

References

[1] Lim, Bee, et al. "Enhanced deep residual networks for single image super-resolution." Proceedings of the IEEE conference on com-

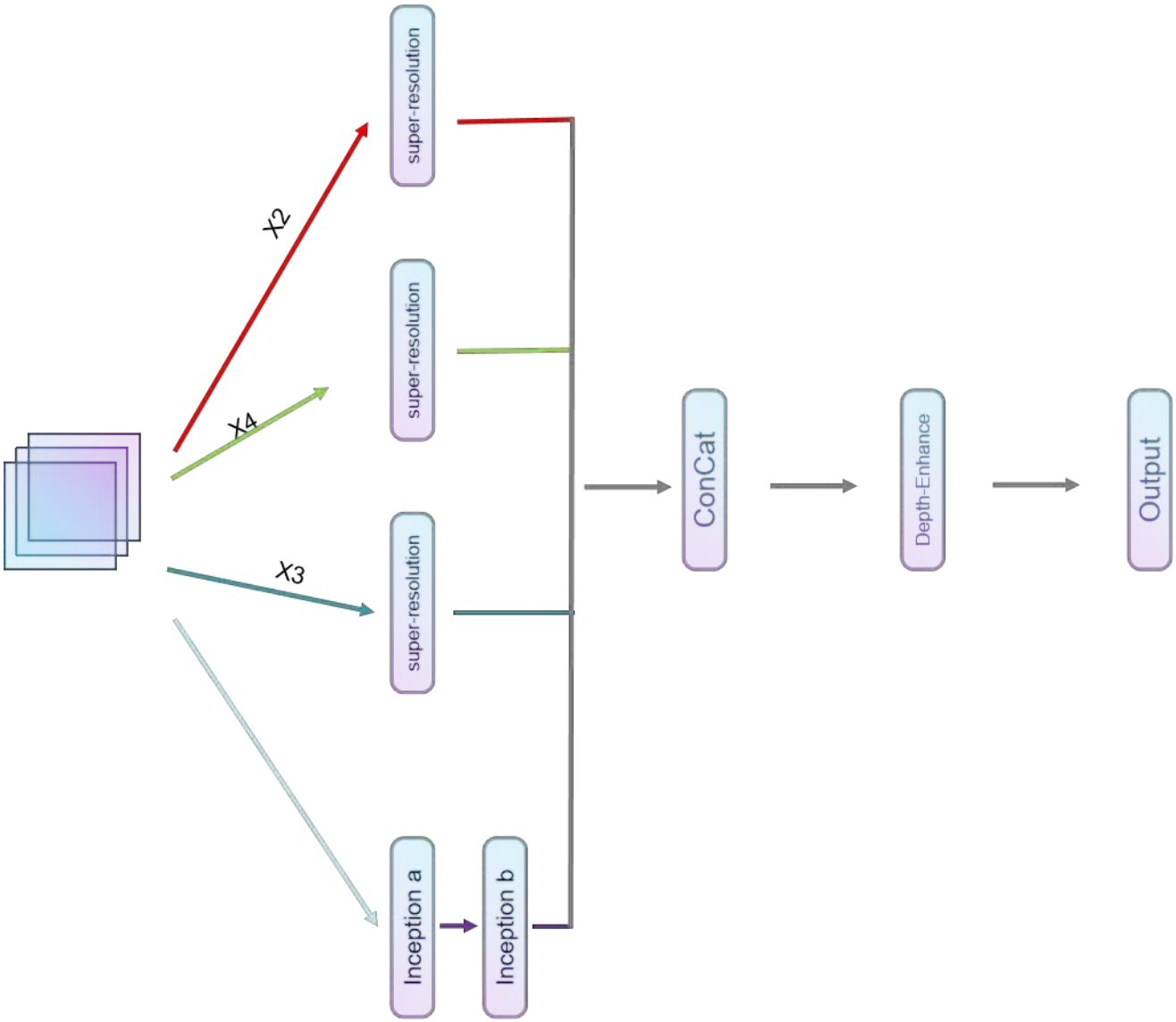


Figure 1: Model

	ESPL		Test		Sintel		Kodak	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
D4 (baseline)	37.7865	0.937	38.2983	0.9644	38.7648	0.9723	37.6039	0.9651
D4 + Inception	39.3811	0.946	39.5382	0.9541	40.4969	0.975	39.3689	0.9719
D4 + SR	39.3093	**0.9527**	39.6588	0.9691	40.5281	0.9767	39.4155	0.972
D4 + Inception + SR	**39.4002**	0.9462	**40.102**	**0.9711**	**40.892**	**0.9783**	**39.697**	**0.9714**

Table 1: Preliminary results of our approach for the 4 bits to 8 bits case. We can see how our approach outperforms the baseline method.

puter vision and pattern recognition workshops. 2017.

[2] Punnappurath, Abhijith, and Michael S. Brown. ”A little bit more: Bitplane-wise bit-depth recovery.” IEEE Transactions on Pattern Analysis and Machine Intelligence 44.12 (2021): 9718-9724.

Xuanshuo Fu I obtained a Bachelor of Science degree in Computer Science and Technology from Northeast Normal University in 2019, and a Master of Engineering in Network Security and Technology from Harbin Engineering University in 2023.

Implicit Neural Representation Model for Camera Relocalization with Learning in Global Multi-Scenes Scenario

Shun Yao
Advisor: G. Mozerov, Yongmei Cheng
Computer Vision Center & Northwestern Polytechnical University
E-mail: syao@cvc.uab.cat

Keywords: Visual relocalization, Scene coordinate prediction, Conditional adaption, Implicit neural representation

1 Summary of Previous and Current Work

In the last two years, I have studied the representation precision problem caused by the discretization of existing digital elevation models. In the meanwhile, different applications require specific discrete representations, which need format conversions. However, these conversions inevitably compromise the fidelity of elevation data. To solve the above problems that lead to inaccurate representation of elevation data, we develop a new continuous representation model (CDEM) that allows height values to be obtained at any arbitrary query position, aiming to preserve the continuity of topographic elevation data in the real world. This work was published in [1].

My recent work focused on visual relocalization across multiple scenes. One critical task in remote sensing is determining the position of a video camera relative to the scene depicted in a series of images captured by the camera. Classical approaches often necessitate pre-built scenario representations, typically obtained using structure-from-motion (SfM) techniques or consisting of posed landmark images, and the implementation of complex algorithms such as image retrieving, feature matching, and absolute/relative pose estimation. In practice, using pre-built scenario representations needs significant storage, poses a risk of exposing private information present in the scene, and is time-consuming for accurate construction. To avoid using pre-built representations, recent methods utilizing scene coordinate (SC) regression-based models to directly predict 3D points in the scene's world coordinates for 2D pixels in query images. A PnP solver within a RANSAC loop is then used to calculate the final camera pose. Figure 1 shows a visual relocalization pipeline by us-

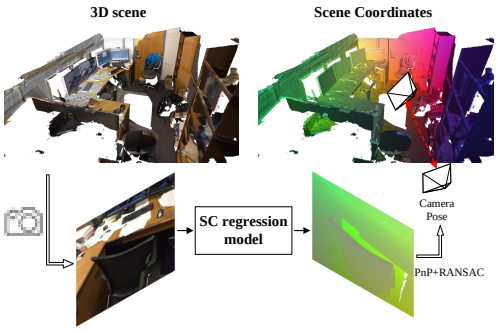


Figure 1: A typical pipeline of using an SC regression model for visual relocalization. We visualize 3D scene coordinates by mapping XYZ to the RGB cube.

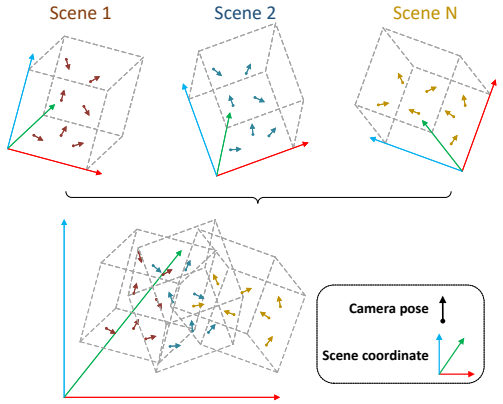


Figure 2: In learning a single SC regression model for visual relocalization across multiple scenes, all scenes are represented in a unified global coordinate system. We color-code the scene frame coordinates to illustrate the differences between individual and joint scene representations.

ing an SC regression model.

These SC regression models have demonstrated promising performance in visual relocalization regarding accuracy and efficiency for a single scene. However, extending SC regression models to multiple scenes typically requires retraining model parameters or constructing reference landmarks, which is a time-consuming process. Instead of separately learning multiple sets of model parameters for different scenes, we use data from all scenes for training. Thus, we represent all scene coordinates within a single global reference coordinate system. Note that the coordinates of different scenes may overlap or be widely distributed. This is illustrated in Figure 2.

Since multiple scenes may have similar image patches corresponding to different scene coordinates, it is challenging to distinguish which scene the patches belong to and to achieve accurate SC predictions using a single regression model. Additionally, different image patches from overlapping scenes may correspond to the same scene coordinate, further complicating model predictions. To address these issues, we propose encoding scene information as a condition and introducing this condition into our model to resolve visual ambiguities across scenes. A key insight is to assign a specific data embedding for each target scene. These embeddings are input into our model for scene coordinate prediction and are optimized during training. The overview of our model is illustrated in Figure 3.

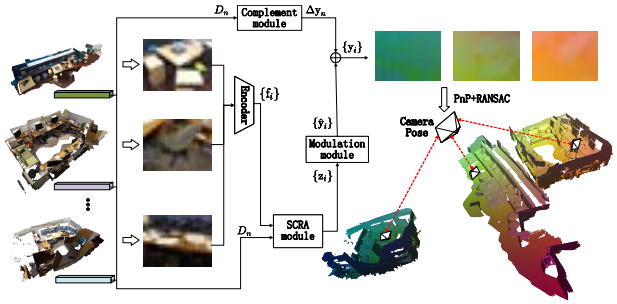


Figure 3: Framework Overview.

2 Future Work and Challenges

The relocalization results of our SCINR are not as precise as those of separately trained ACE models. This is expected, as integrating multiple scenes into a global coordinate complicates SC regression for a single neural model. We believe future research will enhance multi-scene visual relocalization to reduce this accuracy gap. Thus, we consider two directions for future work: first, addressing the imbalance in the number of training samples from different scenes by designing a new loss function or employing a data rebalancing strategy may improve model effectiveness. Second, the calculated loss value of the query image may vary significantly depending on the sampled scene. Training the model with samples from multiple scenes may cause the learning process to focus on scenes with higher loss values. Ensuring the model learns equally from different scene samples may lead to more precise results.

Publications

[1] Yao, Shun, Fei Yang, Yongmei Cheng, and Mikhail G. Mozerov, 3d shapes local geometry codes learning with sdf. In *ICCV Workshop*, pp. 2110-2117, 2021.

[2] Yao, Shun, Yongmei Cheng, Fei Yang, and Mikhail G. Mozerov, A Continuous Digitial Elevation Representation Model for DEM Superresolution. In *ISPRS Journal of Photogrammetry and Remote Sensing*, 208, 1-13, 2024.

[3] Yao, Shun, Yongmei Cheng, Fei Yang, and Mikhail G. Mozerov, Implicit Neural Representation Model for Camera Relocalization with Learning in Global Multi-Scenes Scenario. (Under review of ISPRS JPRS)

Shun Yao received his B.S. degree in Automation from Northwestern Polytechnical University, Xi'an, China, in 2015. He received his M.S. degree in 2017 and is currently a Ph.D. Candidate in the Department of Automation at Northwestern Polytechnical University and computer vision at Universitat Autnoma de Barcelona. His research interests are in the areas of 3D shape representation and deformation, graph convolution, and neural implicit models.

SR-AerIA: Super-Resolution for Aerial Imagery Application

Mohamed Ramzy Ibrahim
Advisor: Robert Benavente, Felipe Lumbreras, Daniel Ponsa
Computer Vision Center & Universitat Autònoma de Barcelona
E-mail: mramzy@cvc.uab.es

Keywords: super-resolution, deep learning, remote sensing, classification, segmentation

1 Summary of Previous and Current Work

In recent decades, various techniques for image super-resolution (SR) have emerged, allowing the generation of high-resolution (HR) images from low-resolution (LR) ones. A key application of SR is in remote sensing, where it addresses challenges such as hardware costs, sensor limitations, and limited bandwidth for transmitting HR images from satellites or UAVs to Earth. SR techniques improve image quality from various sensors, enabling more detailed analysis, and are divided into single-image super-resolution (SISR) and multiple-image super-resolution (MISR).

SISR generates one HR image from one LR image. Initially, traditional approaches relied on learning, interpolation, and optimization techniques, but DL methods now offer superior restoration quality. SISR models struggle with limited information from LR images. In remote sensing, atmospheric conditions or sensor noise can degrade SISR performance. However, SISR is less computationally complex and does not require extensive preprocessing like registration and image alignment.

MISR generates a HR image from multiple LR images. Literature includes spatial, frequency, and deep learning models. Most SoTA models focus on real-time RGB videos, neglecting spectral bands with different frequencies. Models for remote sensing datasets with distinct spectral bands also struggle to utilize image information, such as correlations between multiple LR images and local/global information fusion.

In our first work [1], we propose a 3D deep learning model for MISR (shown in Fig. 1.A). It addresses gaps in SoTA by using 3D convolution layers to capture correlations between multiple LR images and embedding these in dense blocks to enhance information flow. Additionally, the model integrates local and global residual connections to prevent information loss and vanishing gradient issues.

In our second work [2], we propose a model that fuses global information from a multi-headed self-attention (MSA) transformer with local information from a convolutional neural network (RRDB) (shown in Fig. 1.B). This fusion addresses the limitations of each technique: the Transformer modules mitigate the locality of convolutional kernels, while the RRDB addresses issues of patch division, such as border pixel miscommunication and artifacts.

In our third work [3], we investigated the impact of SR techniques on higher tasks like scene classification, examining the correlation between SR metrics (PSNR and SSIM) and classification results (accuracy, precision, and F1-score) (shown in Fig. 1.C). We found that classification results are consistently better with HR and super-resolved images compared to LR images. However, there is variation between SR and classification results; at very low resolutions, the SR models with the best SR metrics do not necessarily yield the best classification outcomes.

Currently, we are working on: first, developing a new SR model and testing its effect on segmentation. Second, we are developing the 3DRRDB model [1] to improve MISR results. We are working on enhancing the 3DRRDB block to focus more on the high-frequency components in the images. Moreover, we are testing different fusion scenarios to show their effects on the results.

2 Future Work and Challenges

First, our ViT-based model could be adapted and tested in various applications and domains, including MISR. Second, the impact of SR on scene classi-

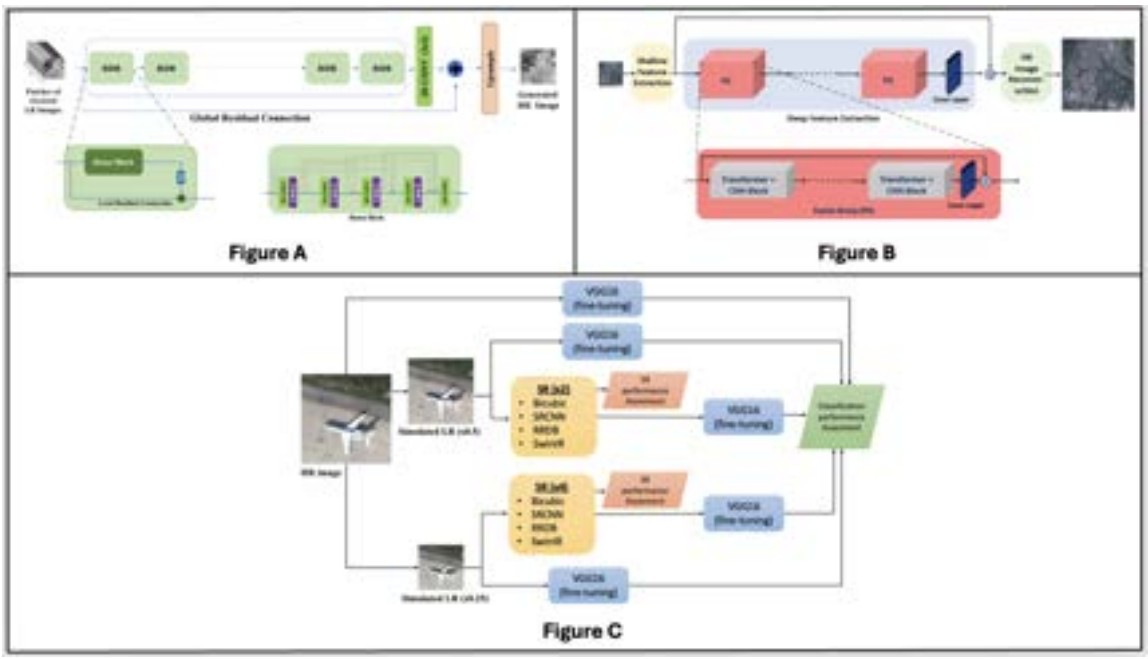


Figure 1: Summary of proposed models. A: The pipline of 3DRDDB model for MISR remote sensing, B: The pipeline of SWViT-RRDB that fuses MSA with RRDB, and C: Diagram for experiments done for showing the effect of SR models on remote sensing scene classification.

fication could be explored further to assess its effect on object detection and image segmentation at very small scaling factors ($\times 8$ and $\times 16$). Finally, we plan to develop a transformer that combines spatial and temporal information to generate HR images from multi-image, multi-spectral LR satellite images.

Publications

[1] M. R. Ibrahim, R. Benavente, F. Lumbreras and D. Ponsa, "3DRRDB: Super Resolution of Multiple Remote Sensing Images using 3D Residual in Residual Dense Blocks," 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, New Orleans, LA, USA, 2022, pp. 322-331, doi: 10.1109/CVPRW56347.2022.00047.

[2] M. R. Ibrahim, R. Benavente, D. Ponsa and F. Lumbreras, "SWViT-RRDB: Shifted Window Vision Transformer Integrating Residual in Residual Dense Block for Remote Sensing Super-Resolution," 2024 *In Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 3: VISAPP; ISBN 978-989-758-679-8, SciTePress, pages 575-582.* DOI: 10.5220/0012399300003660

[3] M. R. Ibrahim, R. Benavente, D. Ponsa and F. Lumbreras, "Unveiling the Influence of Image Super-Resolution on Aerial Scene Classification," 2023 *In: Vasconcelos, V., Domingues, I., Paredes, S. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2023. Lecture Notes in Computer Science, vol 14469. Springer, Cham.* https://doi.org/10.1007/978-3-031-49018-7_16

Mohamed Ramzy Ibrahim received his B.S. and M.S. degrees in Computer Engineering from AASTMT, Alexandria, Egypt, in 2018 and 2020. He is currently a Ph.D. candidate at the CVC-UAB, Barceolna Spain, and an Assistant Lecturer in the Computer Engineering Department at AASTMT. His research interests include Image Processing, Computer Vision, Machine Learning, and Deep Learning.

Color in Visual-Language Models: CLIP deficiencies

Guillem Arias
Advisor:Maria Vanrell
Computer Vision Center
E-mail: guillem.arias@cvc.uab.es

Keywords: Network Visualization, Neuron Feature, Selectivity Index, CLIP, Multimodal Neurons

1 Summary of Previous and Current Work

The goal of this thesis is to create a set of tools to visualize and dissect networks, in order to better understand and improve them. In order to do so, we will be working with the Neuron Feature, a visualization technique based on the weighted average of the images that maximally activate a neuron and Selectivity Index to quantify them As descrived in Understanding trained CNNs by indexing neuron selectivity" in Pattern Recognition Letters.[1], where we also did an extensive review of all the state of the art techniques to visualize and understand Neural Networks, and we provided a taxonomy of methods according to their properties.

During this year we have explored how color is dealt in CLIP (Contrastive Language-Image Pretraining) wich is currently the most influential VML (Visual Language model) in Artificial Intelligence. After performing different experiments on synthetic datasets created for this task, we conclude that CLIP is able to attribute correct color labels to colored visual stimulus. However, we come across two main deficiencies: (a) a clear bias on achromatic stimuli that are poorly related to the color concept, thus white, gray and black are rarely assigned as color labels; and (b) the tendency to use text over other visual information which was previously proved on other tasks, which we proved to be highly significant in color labelling through an exhaustive Stroop test.

With the aim to find the causes of these color deficiencies, we analyse the internal representation at the neuron level. We conclude that CLIP presents an important amount of neurons selective to text and colored text, but a smaller amount of multi-modal color-

Color of the Font		Written Color		Background Color		None
FONT	0.00	White	1.73		0.01	0.00
FONT	0.00	Black	5.74		0.13	0.00
FONT	0.01	Grey	3.05		1.24	0.05
FONT	0.12	Red	4.71		3.50	0.01
FONT	0.40	Green	5.67		3.07	0.01
FONT	0.26	Blue	4.91		3.42	0.00
FONT	0.27	Yellow	6.13		7.74	0.01
FONT	0.00	Brown	6.71		3.06	0.00
FONT	0.26	Orange	6.65		4.80	0.00
FONT	0.21	Pink	6.26		3.14	0.00
FONT	0.83	Purple	7.96		7.92	0.00
TOTAL	2.35%		59.53%		38.03%	0.09%

Figure 1: Experiment 4. Stroop test with Colored Background. *(Column 2): % of Correct Answers. (Columns 4,6,7): % of Incorrect answers.*

text neurons only in middle layers that could help in overcoming these difficulties.

This multi-modal neurons are neurons that encapsulate the "idea" of a color, being activated by any of its modalities: written, or as a real color. And its appearance in shallow layers may signal that although the network learned the basic colors, that information was buried by more complex features more useful in high complexity tasks.

Our investigation also underscores the necessity of refining color representation mechanisms in neural networks to foster a more comprehensive comprehension of colors as humans understand them, thereby advancing the efficacy and versatility of multimodal models like CLIP in real-world scenarios.

2 Future Work and Challenges

Previous studies on Clip understanding have pointed out the existance of multimodal neurons in the last layers of CLIP (close to the image encoder) but the discovery of color Multimodal Neurons in shallow layers let us know that some cimple concepts are understood by the neural network in earlier stages of it,

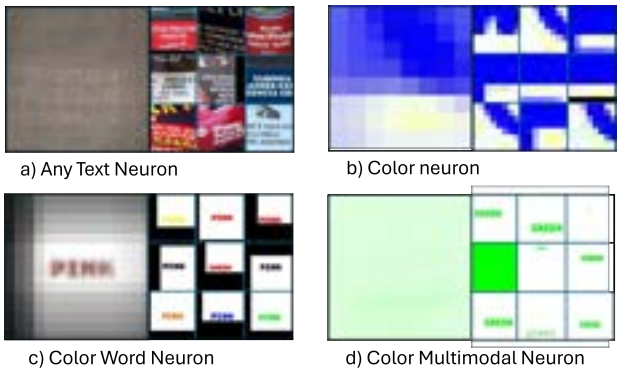


Figure 2: Example of 4 types of Neurons relevant in the Stroop task. Left side: Neuron feature (weighted averaged of the first 100 top-scoring cropped images). Right side: 9 top-scoring cropped images from the Stroop dataset.

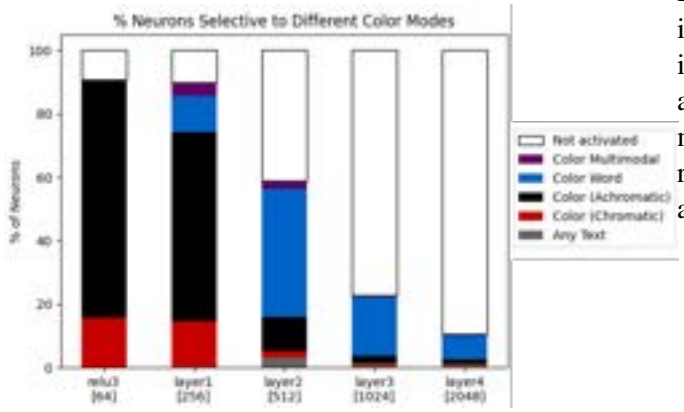


Figure 3: Distribution of Neuron Types per layers in CLIP.

therefore, or next lines of investigation will be:

A) Try to find other multi-modal neurons in shallow layers: Explore if other concepts such as basic shapes or textures are understood earlier in CLIP.

B) Try to use this knowkledge to simplify a lot CLIP to perform simple task: we hipothecise that this multimodal neurons encapsulate all the knowledge required to solve any color related questions, therefore, with an attention mechanism we can lead CLIP to skip all other unnecessary layers to solve those questions.

Publications

[1] Ivet Rafegas and Maria Vanrell and Luis A. Alexandre and Guillem Arias,Understanding trained CNNs by indexing neuron selectivity. In arXiv, eprint: 1702.00382, 2019

[2] Guillem Arias, Ramon Baldrich, Maria Vanrell. Color in Visual-Language Models: CLIP deficiencies. Journal of Imaging Science and Technology. Under Review.

Guillem Arias received his B.S. degree in Biomedical Engineering from Universitat Pompeu Fabra, Barcelona, in 2015. He received his M.S. degree in Brain and Cognition in 2016 and his M.S. degree in Biomedical Engineering in 2018 and is currently a Ph.D. Candidate in the Computer Science Department at the Universitat Autonoma de Barcelona. His research interests are in the areas of machine learning and human cognition and their combination.

Anomaly Detection for Efficient Lesion Screening in Gastrointestinal Endoscopy

Yael Tudela Barroso
Advisor: Jorge Bernal del Nozal
Computer Vision Center & Universistat Autonoma de Barcelona
E-mail: ytudela@cvc.uab.cat

Keywords: Anomaly detection, Medical imaging, semantic segmentation

1 Summary of Previous and Current Work

In the past years, I have been actively engaged in research focusing on anomaly detection for efficient lesion screening in gastrointestinal endoscopy. My work has primarily concentrated on colorectal cancer (CRC) polyp classification, leveraging anomaly detection within gastrointestinal tract imaging. Initially, my research explored various medical and industrial datasets, but it has since pivoted towards developing an efficient method for polyp semantic segmentation incorporating class information.

One of the significant advancements in my research has been the exploration of semi-supervised learning environments, using other public datasets as "unlabeled" data to enhance the robustness of our models while maintaining class specificity. This shift has posed challenges, particularly due to the data-hungry nature of these approaches, which we are actively addressing.

My recent work has shown promise in improving clinician effectiveness by potentially reducing costs associated with the removal of non-cancerous polyps, focusing on anomaly detection within medical imaging, specifically gastrointestinal tract irregularities. This research has been documented and discussed in several publications, such as the work published in [4].

2 Future Work and Challenges

Future research will focus on finalizing the alignment of features from different image modalities, such as white light (WL) and narrowband imaging (NBI), to increase model robustness. We will resume work on semi-supervised learning approaches to leverage large volumes of unlabeled data.

Developing and validating a prototype of the enhanced anomaly detection system will be a key goal. This involves integrating advanced semi-supervised learning techniques to better utilize unlabeled datasets and improve model accuracy. Furthermore, aligning features from various imaging modalities will be essential in creating a comprehensive system.

The ultimate objective is to create a robust and efficient anomaly detection system that contains the whole pipeline (from detection and tracking to providing a fine-grained classification) that can reliably assist clinicians in the early detection and classification of gastrointestinal tract anomalies, thereby improving patient outcomes.

Publications

[1] **Tudela, Y.**, García-Rodríguez, A., Fernández-Esparrach, G., & Bernal, J. "Improving real-time polyp classification via fine-grained segmentation: A Complete benchmark". In *Transactions on Biomedical Engineering*. (Under review)

[2] **Tudela, Y.**, Majó, M., de la Fuente, N., Galdran, A., Krenzer, A., Puppe, F., YamLahi, A., Tran, T.N., Matuszewski, B.J., Fitzgerald, K., Bian, C., Pan, J., Liu, S., Fernández-Esparrach, G., Histace, A., & Bernal, J. "A complete benchmark for polyp detection, segmentation and classification in colonoscopy images". In *Frontiers in Oncology*. (Under review)

[3] García-Rodríguez, A., **Tudela, Y.**, Córdova, H., Carballal, S., Ordás, I., Moreira, L., Vaquero,

E., Ortiz, O., Rivero, L., Sánchez, F.J., Cuatrecasas, M., Pellisé, M., Bernal, J., & Fernández-Esparrach, G. ”In vivo computer-aided diagnosis of colorectal polyps using white light endoscopy”. In *Endoscopy International Open*, Vol 10, pp. E1201-E1207, 2022.

[4] **Tudela, Y.**, García-Rodríguez, A., Fernández-Esparrach, G., & Bernal, J. ”Towards Fine-Grained Polyp Segmentation and Classification”. In *Lecture Notes in Computer Science*, Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging, CLIP 2023, Vancouver, CANADA, Vol 14242, pp. 32-42, 2023.

Yael Tudela received his B.S. degree in Computer Science from Seoul Universitat Autonoma de Barcelona, Spain, in 2019. He received his M.S. degree in Computer Vision in 2020. He is currently a Ph.D. Candidate in the Computer Vision Center of Universitat Autonoma de Barcelona. His research interests are in the areas of medical imaging and self-supervised learning.

Semantic Segmentation in Multispectral Imaging

Leo Ramos
Advisor: Ángel Sappa, PhD
Computer Vision Center & Universitat Autònoma de Barcelona
E-mails: ltramos@cvc.uab.cat; asappa@cvc.uab.cat

Keywords: semantic segmentation, multispectral imaging, computer vision, deep learning

1 Summary of Previous and Current Work

Semantic Segmentation (SS) is a fundamental task in Computer Vision, involving the partitioning of a digital image into distinct segments or sets of pixels [1]. Unlike simpler forms of image analysis based on color or texture, SS assigns each pixel to predefined classes, enhancing the contextual understanding of the image [2]. This process utilizes algorithms and techniques to parse and classify image elements at a granular level [3]. Consequently, the output is a segmented map where pixels represent meaningful labels corresponding to specific categories or objects.

Traditionally, SS has predominantly been applied to images captured in the RGB range. While effective for many applications, RGB imagery sometimes lacks the detail or information necessary for precise segmentation in certain scenarios and applications [4]. Here, Multispectral Imagery (MSI) emerges as a pivotal alternative, providing a comprehensive view of scenes. MSI involves capturing image data across specific wavelengths of the electromagnetic spectrum [5]. Each spectral band responds uniquely to different properties, enhancing the ability to distinguish among various materials and features.

In the past six months, which encompass the time dedicated to developing this thesis since its official start, our primary focus has been to delve into the intricate realm of MSI and SS. This exploration has involved establishing foundational knowledge, understanding the historical context, and identifying key aspects such as datasets, metrics, potential application fields, gaps, and trends for conceptualized contributions that will be further developed in upcoming stages.

As a result, three specific areas of potential application have been identified: Land Cover/Land Use, Agriculture, and Urban Environments. Linked to these applications, several datasets have been identified with multispectral configurations including different spectrums/bands such as Infrared, Thermal, Red Edge, Short-wave Infrared, in addition to RGB. Furthermore, several predominant approaches and methods have been identified in the realm of SS in MSI, which currently serve and will serve as the basis for the development of new methodological proposals planned for implementation. The results of these analyses and reviews are scheduled for publication, and work is underway on writing the articles (see Publications).

2 Future Work and Challenges

In the upcoming academic term, which in my case also includes the latter half of the current one, the plan is to finalize laying the foundations of the research and begin with the most innovative and complex proposals. Specifically, the following objectives are planned:

1. Complete the writing and publication of the reviews and preliminary studies currently in progress.

2. Conceptualize a study for the generation of multispectral synthetic images for semantic segmentation.

3. Evaluate different architectures and methods for generating synthetic images such as GANs and diffusion models.

4. Perform a qualitative and quantitative analysis of synthetic images against real images to assess methods effectiveness.

5. Conduct a comparative analysis of spectral information derived from synthetic images.

6. Augment an RGB dataset with additional synthetic bands and release it.

7. Write a new scientific article and submit it to a conference or journal.

Once these objectives are completed, in the following academic terms, the plan is to continue with other areas of interest such as image fusion, super resolution, and development of new neural network architectures for SS.

Publications

[1] Ramos L., Sappa A., "Multispectral Semantic Segmentation for Land Cover Classification: An Overview." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, (1st revision), 2024.

[2] Ramos L., Sappa A., "Literature Review of Semantic Segmentation in Multispectral Imaging for Agriculture Applications." (under preparation), 2024.

[3] Ramos L., Sappa A., "Performance Comparison of Deep Learning Methods for Land Cover Classification Using Multispectral Imaging." (under preparation), 2024.

References

[1] H. Shijie, Zhou Y., Guo Y. "A brief survey on semantic segmentation with deep learning." *Neurocomputing*, 2020.

[2] Su D., Kong H., Qiao Y., Sukkarieh S. "Data augmentation for deep learning based semantic segmentation and crop-weed classification in agricultural robotics." *Computers and Electronics in Agriculture*, 2021.

[3] Lee J., Kim E., Yoon S. "Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[4] Sun L., Yang K., Hu X., Hu W., Wang K. "Real-time fusion network for rgb-d semantic segmentation incorporating unexpected obstacle detection for road-driving images." *IEEE Robotics and Automation Letters*, 2020.

[5] Noor E. Shaik, Bryce Widdicombe, Dechuan Sun, Sam E. John, Dongryeol Ryu, Am- palavanapillai Nirmalathas, and Ranjith R. Unnithan. "Longwave infrared multispectral image sensor system using aluminum-germanium plasmonic filter arrays." *Nano Research*, 2023.

Leo Ramos received the degree (Hons) in information technology engineering with a focus on artificial intelligence from Yachay Tech University, Ecuador, in 2023. He is currently pursuing the Ph.D. degree in computer science with the Computer Vision Center, Universitat Autónoma de Barcelona, Spain. He is an associate researcher at Kauel Inc., Silicon Valley, CA, USA, leading scientific research in computer vision and intelligent systems for the energy sector. His research interests include deep learning, computer vision, natural language processing, and industry applications. In line with these interests, he collaborates with renowned researchers from Spain, France, U.K., UAE, and the USA, including collaborations with NASA scientists.

Rating symmetry in digital images: a biologically-inspired approach

Mehdi Fatan Serj
Advisor: Carlos Alejandro Parraga
Computer Vision Center & Universitat Autonoma de Barcelona
E-mail: mfatan@cvc.uab.es

1 Abstract

Symmetry fascinates us. Its presence in everything from the detailed patterns of snowflakes to the grand design of the Parthenon has sparked human curiosity for thousands of years. This essential concept goes beyond art and beauty, impacting fields such as architecture, biology, and computer science. This research explores the intriguing world of symmetry, examining its appearances in various areas and its possible link to human perception through the use of artificial neural networks. The focus is on bio-inspired models to discover new phenomena in the perception of symmetry and its role in the lives of different creatures.

Keywords: Symmetry, Brain, Layers, Biology, Artificial Neural Network

2 Summary of Previous and Current Work

Human symmetry perception, though seemingly straightforward, has been extensively studied since 1886, starting with Ernst Mach's *The Analysis of Sensations* [15]. Mach noted the prominence of vertical axis symmetry, which later research confirmed as humans detect bilateral symmetry more quickly and easily than other forms such as rotational or translational symmetry [40]. Further studies suggest that the human visual system can compute symmetry automatically, pre-attentively, and effortlessly, detecting it globally at a first glance [1]. Infants as young as four months can discern different types of symmetry around an axis [16].

2.1 Early Visual Processing

In the early visual system, neurons filter images at different scales before computing symmetry [17]. Symmetry perception aids in the perceptual grouping of contour fragments through parallel filtering operations [14]. Various mechanisms, including texture processing [2] and color processing [5], are involved in detecting symmetry in human vision. Symmetry perception may involve two operations, similar to object detection neural networks: rapid detection in multiple orientations followed by detailed evaluation [39]. The brain's need to recognize objects regardless of their position and orientation may drive symmetry computation [12].

2.2 Factors Influencing Symmetry Perception

Studies show that low spatial frequency channels contribute more to symmetry perception than high spatial frequency channels. For instance, combining random-dot arrays with vertical and horizontal symmetry enables humans to perceive both forms of symmetry [23]. Symmetry detection is categorized as a set of low-cost operations that analyze an image by dividing it into different filters and parts. However, symmetry perception is influenced by several factors, including the number of perceived objects, depth on a screen, and the orientation of the symmetry axes [11]. The location of the symmetric stimulus is crucial, as symmetry is easier to detect when the axis is at the point of fixation [24]. Perturbations near the center of a symmetric pattern are more noticeable than those on the periphery [22]. Symmetry perception remains robust against blurring, lightening, darkening, and noise [7].

2.3 Pattern Outlines and Symmetry Computation

Pattern outlines may be relevant for symmetry computation, creating the impression of seeing a symmetric object [13]. Experiments outside these outlines, which perturb image shapes, increased reaction time and reduced accuracy [4]. Symmetric patterns are processed faster than asymmetric ones due to requiring less information, reflecting their intrinsic redundancy [6]. Symmetry is a fundamental feature of Gestalt theory, which posits that perception depends on laws like similarity, continuation, closure, proximity, and symmetry [3].

2.4 Dynamic Symmetry and Temporal Aspects

Symmetry perception can be static or dynamic. Static symmetry involves replicating elements across an axis, while dynamic symmetry balances compositional elements around an imaginary axis, enhancing visual exploration [20]. Temporal aspects, such as timeframe symmetry, occur when non-symmetric objects appear symmetric under certain transformations [21]. Research on color in symmetry perception shows contrasting theories: some suggest symmetry mechanisms are color-selective, while others say they are color-sensitive [36]. Chromatic symmetry stimuli result in slower responses than achromatic ones. Attention's role in symmetry perception is also under ongoing research.

2.5 Computational Approaches

Recent advancements have enabled the creation of algorithms and tools inspired by human visual systems to handle 2D and 3D symmetry detection [36]. However, studies indicate that machines may struggle to achieve the same efficiency and speed in symmetry detection as humans [9]. Convolutional neural networks (CNNs) are particularly suited for image and pattern recognition tasks, performing filtering operations similar to the human visual system. Studies show spatial alignment between CNN layers and human fMRI responses, indicating parallel computations [19, 18]. Metrics like the brain hierarchy score quantify the correspondence between deep neural network activations and human brain activity [19]. Despite similarities, neural networks differ from human vision, especially in robustness to image manipulations [38, 11].

Research suggests that artificial neural networks may have inherent symmetry detection capabilities [8]. Neural networks trained on datasets like ImageNet may learn symmetry to differentiate between images [10]. Certain neurons may be dedicated to symmetry computation, similar to mechanisms in early human visual processing [35, 37].

2.6 Biologically Plausible Approaches

This study builds upon a biologically plausible approach initiated by Rolls [26] for transform-invariant object recognition. Subsequent advancements include a full model by Wallis and Rolls [27], and an updated description with significant results by Rolls [28]. Further developments integrate saliency in the dorsal visual system for object fixation and invariant object recognition in the ventral system [34]. The study also explores how VisNet employs slow learning to grasp non-accidental properties of objects [32], and to recognize visually distinct views of objects [25]. The methodology extends to object-based motion representation in the dorsal visual system [33], and to coordinate transforms in the dorsal visual system and parietal cortex for allocentric coordinates [29]. This approach aids in updating hippocampal spatial view cells during navigation [31], showcasing the utility of VisNet's statistical learning from environmental inputs across various brain systems. The paper contrasts unsupervised slow learning in VisNet with current deep learning approaches in vision, emphasizing principles beneficial for future neural network developments [30]. As part of our research, we aim to apply VisNet to learn from a set of generated symmetric objects. This study will explore the model's layers to gain insights into how the human brain perceives symmetry as critical features for object recognition and localization.

3 Future Work and Challenges

Future work in computational models of the visual pathway aims to uncover the fundamentals of symmetry perception. This will involve integrating more realistic biological mechanisms to improve the accuracy of these models. Key areas of focus include detailed cellular and synaptic processes, as well as

the dynamic and adaptable nature of neural circuits. Advances in neuroimaging and electrophysiological techniques will provide richer data, enabling the creation of models that better replicate the hierarchical and parallel processing seen in the visual cortex. Additionally, exploring how multiple senses integrate and the role of feedback connections in visual processing are crucial for understanding symmetry perception.

However, several challenges must be addressed. Highly detailed models require efficient algorithms and high-performance computing due to their complexity and resource demands. The variability and diversity of biological systems make it difficult to create models that are both precise and generalizable. Collaboration across neuroscience, computer science, and cognitive science is essential to develop robust models. Ethical considerations must also be managed carefully, especially regarding the use of AI and machine learning in brain function research. Addressing these challenges will require continuous innovation in both theoretical and applied research, leveraging advancements across various disciplines to uncover the principles of symmetry perception in the visual pathway.

Publications

[1] Symmetry detection by categorization of spatial phase, a model. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 263(1366):105–110, 1996.

[2] Perception, symmetry of art discussed at brain lecture (03/22/02), 2007. Accessed on 06/20/2021.

[3] Ilhom Atadjanov and Sanghoon Lee. Robustness of reflection symmetry detection methods on visual stresses in human perception perspective. *IEEE Access*, 6:63712–63725, October 2018.

[4] H. B. Barlow and B. C. Reeves. The versatility and absolute efficiency of detecting mirror symmetry in random dot displays. *Vision Res.*, 19(7):783–793, 1979.

[5] Anselm Brachmann and Christoph Redies. Evaluating symmetry in images using convolutional neural networks. *Symmetry*, 8:115, 2016. Symmetry assessment using CNNs.

[6] Vicki G. Bruce and Michael J. Morgan. Violations of symmetry and repetition in visual patterns. *Perception*, 4(3):239–249, 1975.

[7] Magnus Enquist and András Arak. Symmetry, beauty and evolution. *Nature*, 372(6502):169–172, 1994.

[8] C. Funk, S. Lee, M. R. Oswald, and et al. 2017 iccv challenge: Detecting symmetry in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1692–1701, 2017.

[9] C. Funk and Y. Liu. Symmetry recaptcha. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[10] C. Funk and Y. Liu. Beyond planar symmetry: Modeling human perception of reflection and rotation symmetries in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 793–803, 2017.

[11] R. Geirhos, D. H. Janssen, H. H. Schütt, J. Rauber, M. Bethge, and F. A. Wichmann. Comparing deep neural networks against humans: Object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*, 2017.

[12] V. Giannouli. Visual symmetry perception. *ENCEPHALOS*, 50:31–42, 201.

[13] Bela Julesz and J. j. Chang. Symmetry perception and spatial-frequency channels. *Perception*, 8(6):711–718, 1979.

[14] P. Locher and C. Nodine. The perceptual value of symmetry. *Computers & Mathematics with Applications*, 17(4):475–484, 1989.

[15] E. Mach. *The Analysis of Sensations*. Dover Publications, 1959.

[16] A. D. Makin, A. Pecchinenda, and M. Bertamini. Implicit affective evaluation of visual symmetry. *Emotion (Washington, D.C.)*, 12(5):1021–1030, 2012.

[17] A. D. Makin, M. M. Wilton, A. Pecchinenda, and M. Bertamini. Symmetry perception and affective responses: A combined eeg/emg study. *Neuropsychologia*, 50(14):3250–3261, 2012.

[18] Y. Mohsenzadeh, C. Mullin, B. Lahner, and et al. Emergence of visual center-periphery spatial organization in deep convolutional neural networks. *Sci Rep*, 10:4638, 2020.

[19] S. Nonaka, K. Majima, S. C. Aoki, and Y. Kamitani. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience*, 24(9):103013, 2021.

[20] W. P. The role of pattern outline in bilateral symmetry detection with briefly flashed dot patterns. *Spatial Vision*, 9(1):57–77, 1995.

[21] W. P. The effects of the contrast polarity of dot-pair partners on the detection of bilateral symmetry. *Perception*, 25(7):757–771, 1996.

[22] Stephen E. Palmer and Kevin Hemenway. Orientation and symmetry: Effects of multiple, rotational, and near symmetries. *J. Exp. Psychol. Hum. Percept. Perform.*, 4(4):691–702, 1978.

[23] S. J. M. Rainville and F. A. A. Kingdom. Spatial-scale contribution to the detection of mirror symmetry in fractal noise. *Journal of the Optical Society of America A*, 16(9):2112, September 1999.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[25] S. Robinson and E. T. Rolls. Recognition of visually different views of objects by visnet. *Neural Networks*, 68:162–173, 2015.

[26] E. T. Rolls. Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. *Hum. Neurobiol.*, 3:209–222, 1984.

[27] E. T. Rolls. Biologically plausible approach to transform-invariant object recognition. *Journal of Neuroscience Methods*, 45(1-2):105–110, 1992.

[28] E. T. Rolls. Updated description with many results on object recognition in the visual system. *Journal of Cognitive Neuroscience*, 24(8):19066–19076, 2012.

[29] E. T. Rolls. Coordinate transforms in the dorsal visual system and parietal cortex. *Journal of Neurophysiology*, 123(5):1789–1801, 2020.

[30] E. T. Rolls. Contrasts between unsupervised slow learning in visnet and deep learning approaches. *Frontiers in Computational Neuroscience*, 15:686239, 2021.

[31] E. T. Rolls. Updating hippocampal spatial view cells during navigation. *Hippocampus*, 31(12):2890–2902, 2021.

[32] E. T. Rolls and L. S. Mills. Learning non-accidental properties of objects in visnet. *Connection Science*, 30(4):578–591, 2018.

[33] E. T. Rolls and S. M. Stringer. Object-based motion in the dorsal visual system. *Trends in Cognitive Sciences*, 10(7):302–307, 2006.

[34] E. T. Rolls and T. J. Webb. Finding and recognizing objects in natural scenes using saliency and invariant object recognition. *Frontiers in Psychology*, 5:123, 2014.

[35] A. Seo, B. Kim, S. Kwak, and M. Cho. Reflection and rotation symmetry detection via equivariant learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9539–9548, 2022.

[36] T. Suk and J. Flusser. Recognition of symmetric 3D bodies. *Symmetry*, 6(3):722–757, 2014.

[37] S. Sundaram, D. Sinha, M. Groth, T. Sasaki, and X. Boix. Recurrent connections facilitate symmetry perception in deep networks. *Sci Rep*, 12:20931, 2022.

[38] L. Tang and D. Ley. Degraded polygons raise fundamental questions of neural network perception. *arXiv preprint arXiv:2306.04955*, 2023.

[39] Matthias S. Treder. Perceptual and neural mechanisms of symmetry detection. *Symmetry*, 2:1510–1543, 2010. Review of psychophysical experiments on symmetry perception.

[40] C. Tyler. *Human Symmetry Perception and Its Computational Analysis*. VSP, 1996.

Mehdi Fatan Serj received his B.S. degree in Computer Engineering from Azad University, Shabestar, Iran, in 2008, and his M.S. degree in Mechatronics Engineering from Azad University, Qazvin, Iran, in 2012. He is currently a Ph.D. candidate in the Department of Computer Vision Center (CVC) at the Universitat Autònoma de Barcelona. His research interests include computer vision, artificial neural networks, bio-inspired algorithms, and computational models of the brain.

Leveraging Color Names for Image Enhancement

David Serrano-Lozano
Advisors: Javier Vazquez-Corral, Luis Herranz
Computer Vision Center & Universitat Autònoma de Barcelona & Universidad Autónoma de Madrid
E-mail: dserrano@cvc.uab.cat

Keywords: Color Naming, Image Enhancement

1 Summary of Previous and Current Work

During the first year of my PhD I have been actively working in the field of image enhancement. Specifically, I have developed a method that leverages Color Naming for retouching images. Our method learns to globally adjust the image for each specific named color via tone curves and then combines the images using an attention-based fusion mechanism to mimic spatial editing. This work has been accepted to the 18th European Conference on Computer Vision 2024.

Image enhancement is crucial in the photo-finishing step of the Image Signal Processor to obtain images that evoke emotions, and enrich storytelling. Despite significant advancements in camera technology, both amateurs and professionals still often resort to post-capture image retouching to enhance an image's quality.

A potential solution to avoid manual adjustment is to learn a deep network model that can mimic the image editing style of a skilled photographer or colorist. These methods leverage a dataset of image pairs with the original and corresponding artist-edited images. It is interesting to consider the tools provided to the artists for performing the image editing. Many photo editing software applications (e.g., Adobe Photoshop[1]) provide users with the ability to manipulate the image based on a small set of fixed colors (e.g., red, green, yellow, orange, blue, purple). Interestingly, the predefined colors selected by software tools are similar to those linguists found to be universal across languages [2], a research topic often referred to as Color Naming.

Our method aims to enhance a low-quality RGB input image x, by a learned model that outputs an enhanced version $\hat{y}$. This enhanced image is derived as close as possible to the expert-retouched image y, based on some objective function L. The approach first applies a DNN backbone that standardizes the input image into a canonical latent space. Next, we use color naming to decompose the image into six color maps. After color naming decomposition, a neural network learns a set of Bezier tone curves to manipulate each color map globally. Finally, an attention mechanism is used to fuse the edited images to achieve local editing effects. Figure **??** shows an overview of our proposed method.

We compared our method with state-of-the-art (SOTA) methods using the widely used MIT-Adobe-5K dataset [3] and the PPR10K dataset [4]. The MIT5K dataset consists of 5000 images captured independently using several DSLR cameras and retouched by five artists. PPR10K is a portrait photo retouching dataset with 11616 high-quality images retouched by 3 experts independently. Our full architecture outperforms contemporary curve estimation, image-to-image, and LUT-based methods across the two datasets.

2 Future Work and Challenges

Regarding the work submitted to ECCV 2024, one challenge we have faced is efficiency. To address this, we plan to improve the architecture by using a less computationally expensive fusion method. We also plan to expand the scope of our method to multi-exposure correction and tone mapping. Color Naming can also be considered a perceptual feature. We also plan to explore how to guide models with both semantic and perceptual features to improve image restoration and enhancement tasks.

During this year, I have also worked on multi-illuminant white balance and we plan to continue exploring this branch.

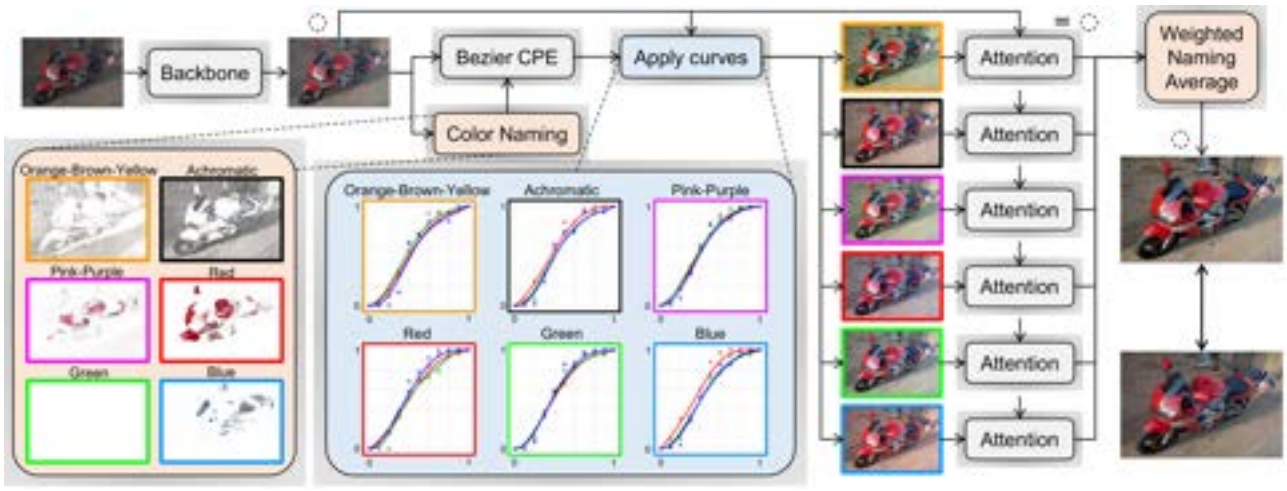


Figure 1: An overview of the proposed method. Our method aims to enhance an input image x. First, we use a UNet-like backbone to standardize the input image into a canonical latent space. Next, we decompose the standardized image $\hat{y}_b$ into six color probability maps (these are shown color-coded in the figure to aid visualization). Next, we learn a set of Bezier curves for each color name to manipulate the standardized image $\hat{y}_b$, obtaining six distinct globally adjusted images. Finally, an attention mechanism is used to fuse the edited images using as *Query* the standardized image $\hat{y}_b$ and as *Key* and *Value* the corresponding edited image. Our learning-based method uses an objective function L to compare the expert-retouched image with our final result $\hat{y}$.

Publications

[1] Serrano-Lozano D., Herranz L., Michael S. Brown and Javier Vazquez-Corral, Learned Image Enhancement via Color Naming. In ECCV, 2024.

References

[1] Adobe Systems: Adobe Photoshop. Computer Software, photoshop.adobe.com

[2] Berlin, B., Kay, P., Basic color terms: Their universality and evolution. In Univ of California Press, 1991.

[3] Bychkovsky, V., Paris, S., Chan, E., Durand, F., Learning photographic global tonal adjustment with a database of input/output image pairs. In CVPR, 2011.

[4] Liang, J., Zeng, H., Cui, M., Xie, X., Zhang, L., PPR10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In CVPR, 2021.

David Serrano-Lozano received his B.S. degree in Telecommunications Engineering from Universitat Politècnica de Catalunya, Barcelona, Spain, in 2021. He received his M.S. degree in 2022 and is currently a Ph.D. Candidate in the Computer Vision Center at the Universitat Autònoma de Barcelona. His research interests are in the areas of low-level computer vision, generative models and computational photography.

Multi-Light Intrinsic-Aware Neural Radiance Fields

Yixiong Yang
Advisor: Maria Vanrell
Computer Vision Center
E-mail: yixiong@cvc.uab.cat

Keywords: relighting, intrinsic decomposition, neural radiance fields (NeRF)

1 Summary of Previous and Current Work

Continuing my research from the previous year, I explored relighting and intrinsic decomposition in the field of NeRF and completed a work titled 'Multi-Light Intrinsic-Aware Neural Radiance Fields'.

Current methods for extracting intrinsic image components, such as reflectance and shading, primarily rely on statistical priors. These methods focus mainly on simple synthetic scenes and isolated objects. They struggle to correctly separate intrinsic components and do not perform well on challenging real-world data. To address this issue, we propose MLI-NeRF, which integrates **M**ultiple **L**ight information in **I**ntrinsic-aware **Ne**ural **R**adiance **F**ields. By leveraging scene information provided by different light source positions complementing the multi-view information, we generate pseudo-label images for reflectance and shading to guide intrinsic image decomposition without the need for ground truth data. Our method introduces straightforward supervision for intrinsic component separation and ensures robustness across diverse scene types. We validate our approach on both synthetic and real-world datasets, outperforming existing state-of-the-art methods. Additionally, we demonstrate its applicability to various image editing tasks.

As illustrated in Fig. 1, our method achieves high-quality intrinsic decomposition results (Fig. 1(c)), as well as novel view synthesis and relighting results(Fig. 1(b)). It also enables applications such as reflectance editing, relighting, and shading editing (Fig. 1 (d)). The results presented in Fig. 2 demonstrate that our method outperforms previous methods.

2 Future Work and Challenges

In the next year, I plan to achieve the following tasks and milestones:

1. Completing the doctoral thesis: The following months will be dedicated to preparing for the defense, and addressing any comments or revisions suggested by my committee.

2. Research into diffusion models: Diffusion has become a hot spot in generative models. I plan to explore how diffusion models can be applied to style transfer in 3D scenes.

Publications

[1] (Submitting) Multi-Light Intrinsic-Aware Neural Radiance Fields. Y. Yang, S. Hu, Haoyu Wu, R. Baldrich, D. Samaras, M. Vanrell.

Yixiong Yang received his B.S. degree and M.S. degree of engineering from Northwestern Polytechnical University, Xi'an, China, in 2018. He also received the second M.S. degree about computer vision in 2020 in School of Engineering at Autonomous University of Barcelona. He is currently a Ph.D. Candidate in the Computer Vision Center of Autonomous University of Barcelona. His research interests are in the areas of interpretation of deep learning networks and image-to-image translation.

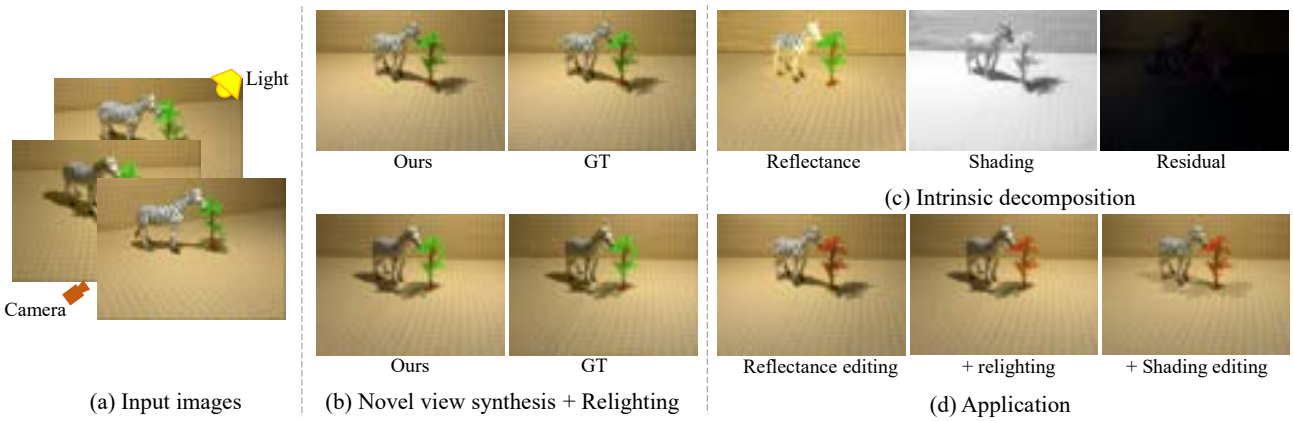


Figure 1: Given real images from ReNe dataset(a), our method learns the neural radiance fields that enable novel view synthesis and relighting (b), and intrinsic decomposition (c) simultaneously. Image editing applications (d), such as reflectance editing, reflectance editing + relighting, and shading editing (simulating two lights), can also be employed.

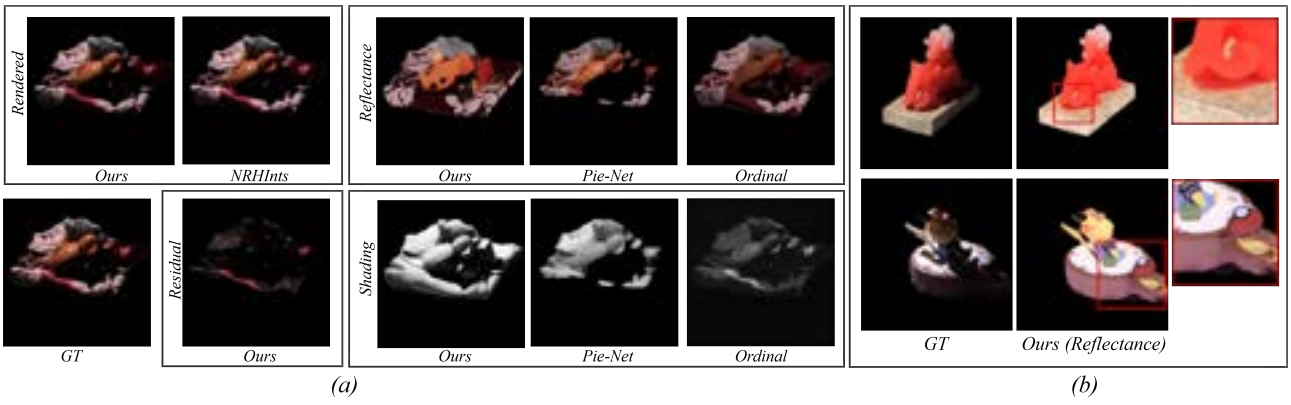


Figure 2: Qualitative Results on the **Real Object Dataset**. (a) Our method compared with NRHints for Rendered map, and Pie-Net and Ordinal for intrinsic decomposition. (b) Our Reflectance estimation for two different scenes, zooming out on object hole and cast shadow area.

Privacy-Aware Document-based VQA

Khanh Nguyen
Advisor: Dimosthenis Karatzas
Computer Vision Center & Universitat Autònoma de Barcelona
E-mail: knguyen@cvc.uab.es

Keywords: federated learning, differetial privacy, docvqa

1 Summary of Previous and Current Work

Document-based VQA (DocVQA), among vision-language (VL) tasks, has made significant advancements through large-scale pre-training on unlabeled datasets. However, existing methods lack privacy guarantees for the sensitive data used to train these models. This requirement is crucial in domains such as finance and healthcare, where data is confidential or personalized, as any leakage of information can pose a serious threat to the linked entity

In [1], we introduce PFL-DocVQA, a large scale DocVQA dataset that mirrors the real-world distribution of invoice documents in different businesses, emphasizing the critical need for safeguarding sensitive information. Our primary aim is to promote research in privacy-aware methods within document analysis. We establish two specific benchmarks: (1) Our first benchmark focuses on developing solutions for training DocVQA models using efficient Federated Learning (FL) algorithms to minimize communication costs per round. (2) The second benchmark extends to integrate DP, aiming to achieve high utility while ensuring a certain level of privacy on the set of all documents from one specific entity. We also highlight privacy issues in state of the art multi-modal LLM models on this dataset, by designing a series of attack exploiting the memorisation effect of the model, and demonstrate its effectiveness in probing different DocVQA models. We explore potential solutions with Differentially Private Federated Learning (DPFL) which satisfy (ϵ, δ)-DP defined at the *group-level*, to protect the identity of invoice owners who contribute their documents to the model's training data. We also evaluate these privacy-aware methods using our proposed attacks, showing that DPFL can achieve a good balance between DocVQA utility and privacy, thereby mitigating privacy issues to a certain degree.

The dataset and benchmark has been announced as parts of the NeurIPS 2023 PFL-DocVQA competition, hosted on the ELSA benchmarks platform[1].

In [2], we explore collaborative training of DocVQA models without sharing training documents between parties. This scenario is particularly important in the document domain due to issues such as legal constraints, copyright, and confidentiality, which limit research in automatic document analysis. We generalize many document understanding tasks in the form of DocVQA, and synthesize a training data distribution by combining different datasets that are carefully selected to ensure a high degree of heterogeneity, reflecting real-world scenarios. We propose a two-stage FL algorithm that includes federated self-training and adaptive server optimization. This algorithm allows our models to effectively learn DU capabilities while controlling gradient variance on the server side, outperforming widely used methods such as FedAvg and FedAdam. Through extensive experiments, we provide valuable empirical insights for further research in this area, demonstrating the impact of several important FL hyperparameters, which may help reduce efforts on the expensive hyperparameter tuning process.

2 Future Work and Challenges

Focusing on privacy-preserving methods for VL tasks, we aim to expand our research in several promising directions: (1) We study behaviors and design attack strategies to highlight the risks associated with a wide range of machine learning models, especially those designed for multimodal data.

[1] https://benchmarks.elsa-ai.eu/

(2) As defenses, we explore potential solutions to effectively train large models for complex tasks such as DocVQA under DP, aiming to achieve a good balance between utility and privacy. (3) We investigate efficient FL methods that enable large models training with FL, as motivated in recent findings, since most of the state-of-the-art DocVQA models are based on Transformers. We also explore possible solutions for unsupervised representation learning in FL scheme, with the goal of replicating the success of pre-training in large scale VL tasks.

Publications

[1] Rubèn Tito, Khanh Nguyen, Marlon Tobaben, Raouf Kerkouche, Mohamed Ali Souibgui, Kangsoo Jung, Lei Kang, Ernest Valveny, Antti Honkela, Mario Fritz, Dimosthenis Karatzas. Privacy-Aware Document Visual Question Answering. In *ICDAR 2024*.

[2] Khanh Nguyen, Dimosthenis Karatzas. Federated Document Visual Question Answering: A Pilot Study. In *ICDAR 2024*.

Khanh Nguyen received his B.S. degree in Electronics Engineering from Hanoi University of Science and Technology in 2018. He received his M.S. degree in Computer Vision in 2021 and is currently a Ph.D. Student advised by Dr. Dimosthenis Karatzas in Computer Vision Center of Universitat Autònoma de Barcelona. His research interest focuses on Privacy, Federated Learning and Vision-Language applications.

Intelligent Reading Systems for Real-World Applications

Sergi Garcia Bordils
Advisor: Dimosthenis Karatzas, Marçal Rusiñol
Computer Vision Center & AllRead MLT
E-mail: sergi.garcia@cvc.uab.es

Keywords: reading systems, text recognition, text detection, deep learning, out-of-vocabulary

1 Previous Work

My work has been mostly centered around text detection and recognition in images (commonly referred to as Scene-Text Recognition or STR). This has been motivated by the necessity of developing models capable of effectively reading text, with specific attention to industrial applications. My work started with the organization of the Out-Of-Vocabulary Challenge[2], a competition on STR that had the objective of assessing the vocabulary relicance of modern STR systems. I also worked on a novel strategy to accelerate transformer-based[1] scene-text detection and recognition models[3]. This strategy removes uninformative tokens at the transformer's input and accelerates training and inference times.

More recently, I introduced the task of Structured Scene-Text Spotting, a task where STR models have to localize and transcribe all the instances of text that follow an arbitrary regular expression (regex). Figure 1 shows our proposed pipeline for this task, the Structured TExt sPotter (STEP), which uses the image and the regex in an end-to-end manner. STEP can spot fixed-length textual patterns that can include spaces, as opposed to the fixed granularity approach of generic models. To assess our approach and compare it to the baselines, we introduced the Structured Scene-Text Spotting Test set, which includes several classes of structured text that follow fixed regular expressions. The experimental results showed how our approach obtains better results compared to state-of-the-art STR methods, even after pairing them with complex post-processing operations.

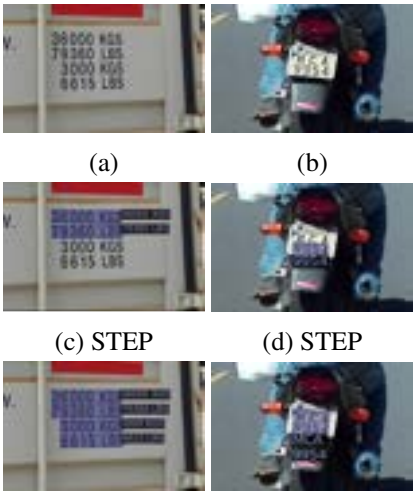
2 Structured Scene-Text Spotting Extension

While STEP was powerful enough to tackle the proposed task, it was limited by the regex operations that it was capable of encoding. Many of the textual patterns that we can find in the real world are variable-length and multi-line patterns (see examples of Figures 2b and 2a), such cases could not be covered by STEP in a single pass. To solve this, I first extended the scope of the Structured Scene-Text Spotting task to cover such cases, thus extending the Structured Scene-Text Spotting Test set with new instances. These new instances present a challenge both to our end-to-end approach and to generic models.

In order to be able to spot these new cases, STEP has developed into STEPup, an iteration of our previous model that is now capable of understanding regex operations that feature variable-length patterns, optional characters, and multi-line text. Figures 2f and 2e show how STEPup can spot the two previous examples that STEP could not cover. Besides the ability to encode new regex operations, STEPup also features an autoregressive text decoder that, as opposed to the parallel decoder of STEP, enforces the recognition output of the network to follow the queried regex, improving the recognition and reducing spelling mistakes.

Once again, the proposed end-to-end approach was compared to generic STR models that were coupled with post-processing operations. The results obtained by STEPup were superior to their generic counterparts and even surpassed the results of STEP in the cases that followed fixed-length patterns, showing the improved decoding capabilities of the proposed conditioned autoregressive text decoder.

Figure 1: STEP takes an image and a regular expression and spots the instances of text of the image that match it.



(a) (b)

(c) STEP (d) STEP

(e) STEPup (f) STEPup

Figure 2

Publications

[1] Sergi Garcia-Bordils, George Tom, Sangeeth Reddy, Minesh Mathew, Marçal Rusiñol, C.V. Jawahar, Dimosthenis Karatzas, Read while you drive - multilingual text tracking on the road. In *Document Analysis Systems*, 2022.

[2] Sergi Garcia-Bordils, Andrés Mafla, Ali Furkan Biten, Oren Nuriel, Aviad Aberdam, Shai Mazor, Ron Litman, and Dimosthenis Karatzas Title. In *European Conference on Computer Vision*, 2022.

[3] Sergi Garcia-Bordils, Dimosthenis Karatzas, Marçal Rusiñol, Accelerating Transformer-Based Scene Text Detection and Recognition via Token Pruning. In *International Conference on Document Analysis and Recognition*, 2023.

[4] George Tom, Minesh Mathew, Sergi Garcia, Dimosthenis Karatzas, C.V. Jawahar, Reading Between the Lanes: Text VideoQA on the Road. In *International Conference on Document Analysis and Recognition*, 2023.

[5] Sergi Garcia-Bordils, Dimosthenis Karatzas, Marçal Rusiñol, STEP: Towards Structured Text Spotting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[6] Sergi Garcia-Bordils, Dimosthenis Karatzas, Marçal Rusiñol, Learning to spot scene-text with regular expressions. Under review in *International Journal of Computer Vision*, 2024.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Sergi Garcia Bordils received his B.S. degree in Computer Engineering from the Autonomous University of Barcelona, Spain, in 2013. He received his M.S. in Computer Vision in 2019 and is currently a research in the Computer Vision Center. His research interests are in the areas of image processing, text detection and recognition and text recognition.

Multi-page document VQA with Recurrent Memory Transformer

Qi Dong
Advisor: Dimosthenis Karatzas, Lei Kang
Computer Vision Center & Universitat Autònoma de Barcelona
E-mail: qdong@cvc.uab.es

Keywords: Document VQA, Multi-page Document VQA , Recurrent Memory Transformer

1 Summary of Previous and Current Work

I have been conducting research in the field of document visual question answering(DocVQA), especially multi-page (DocVQA). Multi-page DocVQA poses a real challenge in the field of document understanding due to its complexity and the amount of information distributed across multiple pages. One of the existing approaches is to treat multiple pages as single-page DocVQA to address the input token limitation. Another approach is to compress the pages to obtain the information of the entire document at once. We propose a new approach to multi-page DocVQA using recurrent memory transformers. This work was published in the 16th IAPR International Workshop On Document Analysis Systems [1]. In our proposed approach,figure 1 shows an overview of the model , pages are processed sequentially while information is passed from one page to the next and enriched with new contents.

2 Future Work and Challenges

We must be responsible for our answers. Existing DocVQA methods lack the ability to produce explanations for their responses. We do not know how the model arrives to the answer. This will reduce people's trust in the model. So we hope to come up with a method that can explain how the answer is obtained.

Our model cannot protect document privacy, and all private content will be exposed to the questioner. So we wanted to come up with a new model. We would like to design a privacy attack DocVQA model to attack private content to curb privacy leaks.

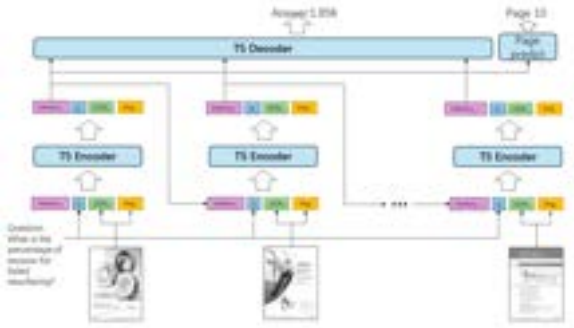


Figure 1: Architecture of the RM-T5 model.

Publications

[1] Qi Dong, Lei Kang, Dimosthenis Karatzas. Multi-page document VQA with Recurrent Memory Transformer. In *Conference*, 16th IAPR International Workshop On Document Analysis Systems(Accepted).

Qi Dong received her B.S. degree in Microelectronics Science and Engineering in 2017 from Hefei University of Technology, Hefei, China. She received her master degree in Computer Technology in 2020 at Xidian University, Xi'an, China. She is currently a Ph.D. Candidate at the Autonomous University of Barcelona in Barcelona, Spain. Her research interests are deep learning, document analysis, DocVQA.

Multi-Modal Structured Reasoning for Historical Document Information Extraction

Adrià Molina Rodríguez
Advisor: Josep Lladós, Oriol Ramos
Computer Vision Center & Universitat Autònoma de Barcelona
E-mail: amolina@cvc.uab.es

Keywords: Graph Reasoning, Historical Documents, Information Retrieval

1 Summary

In the domain of artificial intelligence, graph reasoning involves applying logical structures to a model, allowing it to mimic human cognition processes. For instance, while one wouldn't anticipate discovering a 16th-century document containing a computer purchase receipt, identifying two words linked by a colon as part of the same sentence aligns with logical patterns. While this rules are already learned through pattern recognition in state-of-the-art architectures, they usually require large amount of annotated data and parameters. By introducing such logical structures by design, it is not expected for the model to learn such structure, hence, the quantity of data and required parameters tends to significantly reduce.

The possibility of reducing the quantity of annotated data is particularly interesting in the context of historical document understanding where, given the nature of the data, it is commonly found in sparse languages where distributions differ one to the other even in single-language scenarios where, because of the variance caused due to the large time lapsed within all the documents, language trends and rules significantly changed over time. As shown in Figure 1, globalization and faster communication unified data sources; which serves as a set of common features to take advantage on when it comes to generalisation in document analysis, meaning that **documents share a common language** regardless of its language or, in some case, content. This is not the case in historical sources where, due to input domain differences, it is not usually trivial to generalise from a data perspective.

The document language is encoded through complex relationships between elements of different modalities (images, text, infographics, maths, tables...). While most of the information in the document may be coded through text, the presence of other modalities may affect or contribute to the global understanding of the processed content. In this matter, especially in the context of vision and language, we find several possible interactions:

- **Destructive Relation**: The content of both modalities is not correlated, serving as a distractor for the task. Multi-modality does not provide any meaningful context.
- **Redundant Relation:** The content on both modalities refer to the same topic or expresses the same information. Multi-modality serves as an ensemble mechanism where the architecture may be optimized toward the most trivial solution.
- **Complementary Relation**: The intersection of

Figure 1: Example of different domains in modern (left) and historical (right) document analysis. The differences in the input domain have been narrowed as a consequence of faster communication and common technologies on the production of documents, while documents share a common language, the underlying relations in historical documents are rather sparse.

Figure 2: Fragment of a document showing textual tokens (blue) indicating the presence of an advertisement. The word "typewriter" is deteriorated and truncated (dashed), posing challenges for OCR-based approaches. By identifying the depicted object (red) as a "typewriter" the missing gap in the textual domain can be bridged in a language-agnostic manner.

the content in both modalities is not total, but both contain information relevant for the task. As seen in Figure 2, multi-modality is mandatory in order to fully characterize the content of the document.

In summary, this thesis aims to build a robust historical document understanding framework capable of deciphering the document language through the relationships of the implied elements. As it will be fatherly developed, this is intended to operate from the **contextual**, **structural** and **linguistic** semantics.

2 Future Work and Challenges

The methodology and areas of impact of this research are divided into three levels of information encoding. We refer to the contextual aspects of the document as *external relationships*; i.e., historical context, roles of the named entities appearing in the document, etc. Alternatively, *internal relations* emerge from the document layout itself. This includes the identification of named entities and topics with respect to the appearance of figures and images, the influence of layout in determining the content of the document, or dividing the multifaceted existence of topics within a single page. Lastly, *intrinsic properties* of the document refer to those features that occur at the pixel level and provide a comprehensive view of the document: font, script, and document degradation are typical issues that contribute to the sparsity of historical document language. A framework capable of operating in a wide range of scenarios needs to be robust to domain shifts; therefore, it is crucial to identify the different sets of recognition necessities

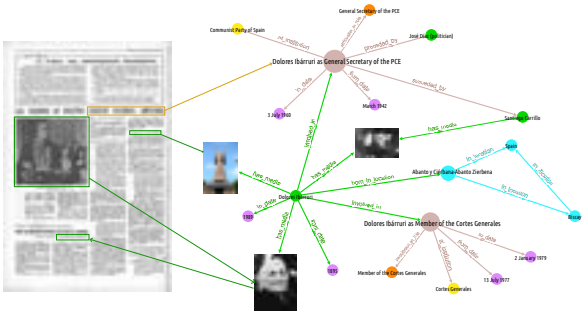


Figure 3: Example of an edge prediction task addressed in this thesis, on which it is expected for the system to capture relationships within a document and an external knowledge base.

that are usually correlated.

2.1 Capturing External Relations

As previously defined, external relationships within a set of multi-modal sources of historical data refer to those connections established from an element of a document to a concept or media registered in a knowledge base, represented in Figure 3. In the context of captioning, some researchers work with knowledge bases from the retrieval perspective [1] or graph indexing [19]. The exploration of graphs for incorporating external knowledge and interpretability via multi-modal message passing represents an intriguing path. As noted in [17], this structural paradigm proves less data-intensive and promotes the interpretability of relationships while allowing control over accessed information. In this vein, the field of scene graph generation [8] stands out, where message passing is a common practice [18, 10].

In the realm of cultural heritage, several datasets have emerged over the years. One notable contribution, dating back to 2019, is MusicKG by Eyharabide *et al.*, presenting a multimodal music knowledge graph containing art depicting musical-related scenes across an eight-century span. Similarly, Candela *et al.* (2022) [2] constructed a multi-modal ontology based on the Mexico Colonial Archive, incorporating maps from various regions produced during the 16th century. In a more recent development, Fan *et al.* (2023) [5] introduced a multi-modal knowledge graph containing historical documents and images related to different arts and cultural aspects of the Chinese intangible cultural heritage. Alternatively,

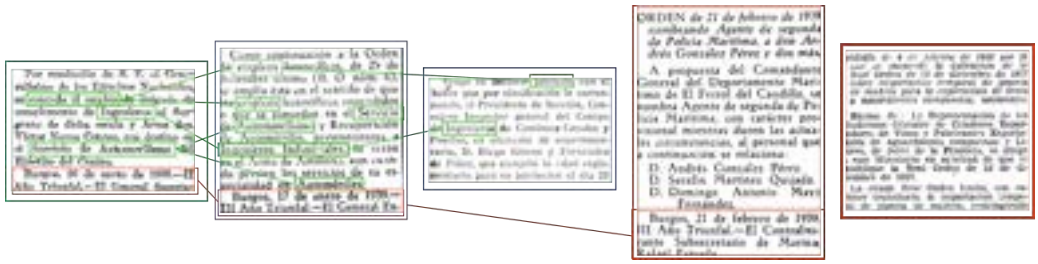


Figure 4: Qualitative example of the **FAS** visual baseline, we observe that given a query consisting on "documents related to engineering job promotion", the top-k retrieved content consist on fragments referring to the desired topic while sharing common visual and date attributes.

and similarly to the aim of the Xarxes project[1], focusing on social structures rather than content, Hyvönen *et al.*[6] (2019) and Debruyne *et al.* [3] (2022) proposed ontologies extracted from documents and biographies from Finland and Ireland, respectively.

2.2 Capturing Internal Relations

Internal relationships within a document are typically expressed through layout constraints. In most documents, one can find a hierarchical variety of topics as the granularity of the recognition procedure deepens. In this context, topics within the same page are likely more related than extra-page content. Upon observing sections in the page, or paragraphs in the sections, the content should also align in the same hierarchical way. Therefore, under this logical assumption, this thesis aims to seek unsupervised methods for encoding topics from the structural properties of documents in a multi-modal setup to alleviate the lack of annotations in historical document analysis.

In that sense, this work expects visual elements like images, tables, or titles to condition the content encoded through the paragraphs of the document. Our work has already produced a benchmark, **FAS** (presented at DAS 2024, *FAS: A Large-Scale OCR-Free Benchmark for Historical Document Retrieval*). This contribution addresses the necessity of evaluating document analysis systems that aim to retrieve content in historical multifaceted documents. However, despite the incorporation of a wide date and layout variety, this work is limited by the absence of multi-modal information within the analyzed pages. Our future work suggests both the incorporation of sources of multi-modal data (historical documents containing images and infographics) and a wider language diversity to ensure a domain-covering evaluation.

During the construction of **FAS**, we realized how the proposed visual baseline tends to retrieve fragments of documents with a strong topic correlation, but also similar layouts and similar dates. This observation (see Figure 4) served as a hint for proving the hypothesis of the existence of a system capable of interpreting topics as an emergent property from structure.

2.3 Capturing Intrinsic Relations

In the most inward aspect of document analysis, we consider the inner relationships within the individual words that compose the documents to be a relevant aspect of this research. Regardless of the semantic content of these words, recognition systems aim to identify the present character strings in the most seamless manner possible, given the limitations of the architectures themselves.

The optimal approach to addressing challenges in pattern recognition involves adopting a "mixture of experts" framework, initially proposed by Jacobs *et al.* (1991) [7]. This framework utilizes multiple specialized models or "experts" to handle diverse and sparse domains within the input data. In deep learning, leveraging a mixture of models has effectively managed domain sparsity. Eigen *et al.* (2013) [4] introduced this concept, concatenating outputs to create a deep representation. Shazeer *et al.* (2017) [16] further demonstrated the capabilities of an outrageously large neural network using the Mixture of Experts (MoE) layer. Notably, MoE models have been integrated into transformers [9], showcasing their scalability and domain-coverage advantages.

[1] http://dag.cvc.uab.es/xarxes/

In this regard, this thesis aims to build recognition systems that can take advantage of the information from different available domains in order to generalize to further unseen languages and scripts. This means that the knowledge generated by the aggregation of scripts should be easily transferable to further data. Once we can cluster the different scripts within a set of documents, we can proceed to exploit recognition systems by aggregating models from similar domains, creating a bootstrapped representation that substantially improves the adoption of new scripts and languages. These results are reported in the NeurIPS 2024 submission "*Towards Domain Shift Immunity: Bootstrapping OCRs with Task Arithmetic*", in which we demonstrate performance improvements of up to three times in the adoption of both non-Latin languages and ciphered scripts compared to classical fine-tuning.

Publications

Publications Related to PhD

- [13], Fetch-A-Set: A Large-Scale OCR-Free Benchmark for Historical Document Retrieval, DAS 2024
- (under review), Towards Domain Shift Immunity: Bootstrapping OCRs with Task Arithmetic, NeurIPS 2024

Previous and Other Publications

- [12], Date estimation in the wild of scanned historical photos: An image retrieval approach, ICDAR 2021
- [11], A Generic Image Retrieval Method for Date Estimation of Historical Document Collections, DAS 2022
- [15], Learning to rank words: Optimizing ranking metrics for word spotting, ICDAR 2021 (non-corresponding)
- [14], A Transformer-Based Object-Centric Approach for Date Estimation of Historical Photographs, ECIR 2024 (non-corresponding)

Adri Molina received his B.S. degree in Data Engineering from Universitat Autònoma de Barcelona in 2022. He received his M.S. degree in 2023 and is currently a Ph.D. Candidate in the Document Analysis Group. He did not receive the Best Student Paper Award from anywhere, he is not trying because he keeps saying "science is no competition". His research interests are in the areas of cultural heritage, data governance, responsible AI, retrieval systems, data visualization and graphs.

References

[1] Bai, Z., Nakashima, Y., Garcia, N.: Explain me the painting: Multi-topic knowledgeable art description generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5422–5432 (October 2021)

[2] Candela, G., Pereda, J., Sáez, D., Escobar, P., Sánchez, A., Torres, A.V., Palacios, A.A., McDonough, K., Murrieta-Flores, P.: An ontological approach for unlocking the colonial archive. ACM Journal on Computing and Cultural Heritage (2022)

[3] Debruyne, C., Munnelly, G., Kilgallon, L., O'Sullivan, D., Crooks, P.: Creating a knowledge graph for ireland's lost history: knowledge engineering and curation in the beyond 2022 project. ACM Journal on Computing and Cultural Heritage (JOCCH) **15**(2), 1–25 (2022)

[4] Eigen, D., Ranzato, M., Sutskever, I.: Learning factored representations in a deep mixture of experts. arXiv preprint arXiv:1312.4314 (2013)

[5] Fan, T., Wang, H., Hodel, T.: Cichmkg: a large-scale and comprehensive chinese intangible cultural heritage multimodal knowledge graph. Heritage Science **11**(1), 1–18 (2023)

[6] Hyvönen, E., Rantala, H.: Knowledge-based relation discovery in cultural heritage knowledge graphs. Digital Humanities in Nordic Countries (2019)

[7] Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural computation **3**(1), 79–87 (1991)

[8] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting

language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017)

[9] Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668 (2020)

[10] Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: Proceedings of the IEEE international conference on computer vision. pp. 1261–1270 (2017)

[11] Molina, A., Gomez, L., Ramos Terrades, O., Lladós, J.: A generic image retrieval method for date estimation of historical document collections. In: International Workshop on Document Analysis Systems. pp. 583–597. Springer (2022)

[12] Molina, A., Riba, P., Gomez, L., Ramos-Terrades, O., Lladós, J.: Date estimation in the wild of scanned historical photos: An image retrieval approach. In: International Conference on Document Analysis and Recognition. pp. 306–320. Springer (2021)

[13] Molina, A., Terrades, O.R., Lladós, J.: Fetch-a-set: A large-scale ocr-free benchmark for historical document retrieval (2024)

[14] Net, F., Hernández, N., Molina, A., Gómez, L.: A transformer-based object-centric approach for date estimation of historical photographs. In: European Conference on Information Retrieval. pp. 137–150. Springer (2024)

[15] Riba, P., Molina, A., Gomez, L., Ramos-Terrades, O., Lladós, J.: Learning to rank words: Optimizing ranking metrics for word spotting. In: International Conference on Document Analysis and Recognition. pp. 381–395. Springer (2021)

[16] Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)

[17] Wang, Y., Yasunaga, M., Ren, H., Wada, S., Leskovec, J.: Vqa-gnn: Reasoning with multimodal semantic graph for visual question answering. arXiv preprint arXiv:2205.11501 (2022)

[18] Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5410–5419 (2017)

[19] Zhao, W., Wu, X.: Boosting entity-aware image captioning with multi-modal knowledge graph. IEEE Transactions on Multimedia (2023)

Multimodal Models for Comics Understanding

Emanuele Vivoli[1,2]
Advisors: Dimosthenis Karatzas[1], Marco Bertini[2]
[1]*Computer Vision Center &* [2]*MICC, University of Florence*
E-mail: evivoli@cvc.uab.es

Keywords: comics, manga, vqa, datasets, tasks, understanding, MLLMs

1 Summary of Previous and Current Work

The initial year was dedicated to a comprehensive review of the comics field, in terms of existing tasks, datasets, and methods. This has led to the writing of a **comprehensive survey on Comics Understanding** [1] which will be submitted to journals by the end of August.

An observed inconsistency in the detection task results across various datasets and difficulty in reproducibility has led to the development of a unified benchmarking system, Comics Dataset Framework [2], which unify several dataset annotations to a unified version that can be easily shared and improved, as shown in Figure 1. We thoroughly benchmark the existing datasets and enrich them with refined annotations manually and through automated processes. The paper has been published at the ICDAR workshop on comics (MANPU).

Additionally, we have noticed a gap between the tasks the SOTA models can operate in comics and the evaluation settings these models go through. In particular, models like Magi, are designed to do multiple tasks together, without any evaluation set to test on. To fill the gap, we have proposed CoMix, a comprehensive benchmark - as shown in Figure 2 - for detection, linking, character naming, and dialog generation [3].

Lastly, with the intent of advancing multimodal models for comics tasks, we have designed a VL-T5 model - as shown in Figure 3 - for the task of text-cloze in comics [4], which has been accepted and will be presented at ICDAR this September.

2 Future Work and Challenges

In the Vision-Language domain, many Multimodal Large Language Models (MLLMs) have been proposed so far, from the more generalist to the one tailored to a specific domain. However, despite many with good OCR capabilities, reading and understanding comic pages remain a limitation. To this end, we are designing a Comic instruction dataset to fine-tune these MLLMs and explore whether understanding comics only requires "seeing" more comics, or, instead, specific architectural and data design.

The next phase involves developing a ComicsVQA dataset accessible to the public and establishing benchmarks tailored for testing (and hopefully also training) multimodal models across single and multi-page comic formats. This also includes innovating pre-training algorithms for data-scarce scenarios like comics. Current progress includes the setup of an annotation website, selection of 100 comic books from COMICS based on their simplicity, language, and open copyright status. Post-annotation, a rigorous evaluation of the annotations and an update of the annotation platform and policies are planned. Explorations into captioning annotations to generate VQA data for single-panel images are also underway.

Another area of exploration is multimodal panel retrieval within comics, a task of considerable interest to comic aficionados that falls under the broader scope of Personalized Image Retrieval. This involves identifying specific panels or sequences where characters engage in particular actions or dialogues. As a novel concept in comics, it introduces the challenge of embedding intricate visual and textual elements in character-specific narratives. This necessitates the development of novel pre-training and fine-tuning strategies to effectively integrate detailed comic text-image pairs.

Focusing on the more challenging aspects of

comic analysis, as listed in the Layers of Comics Understanding [1], our research will expand, at the end of the second year, into innovative domains such as audio generation from panels and conversion of textual narratives into comic strips. Our future research directions include: (1) Proposing small yet high-quality test sets for underexplored areas like VQA and retrieval in comics. (2) Investigating new methods to adapt transformer architectures for comic inputs, both single-page and multi-page. We believe that the unique layout structure of comics is not effectively captured by standard image-patch techniques or more intricate layout representations.

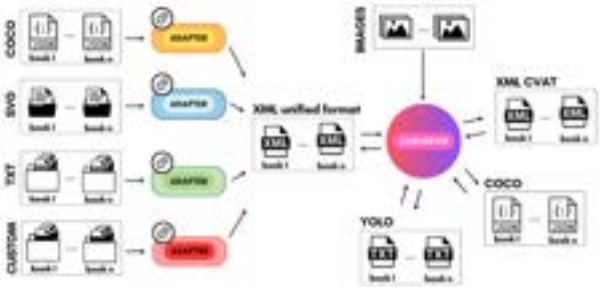


Figure 1: Comics Dataset framework schema

Figure 2: Dataset and task scheme

Publications

[1] Emanuele Vivoli, Andrey Barsky, Marco Bertini, Dimosthenis Karatzas, "The Comics Survey: Datasets, Tasks, and the Future of Comics Understanding", *soon to be submitted*, 2024.

[2] Emanuele Vivoli, Irene Campaioli, Mariateresa Nardooni, Niccolò Biondi, Marco Bertini, Dimosthenis Karatzas, "Comics Datasets Framework: Mix of Comics datasets for detection benchmarking", In *MANPU - ICDAR workshop*, 2024.

[3] Emanuele Vivoli, Marco Bertini, Dimosthenis Karatsaz, "CoMix: A Comprehensive Benchmark for Multi-Task Comic Understanding", *under submission*, 2024.

[4] Emanuele Vivoli, Joan Lafuente Baeza, Ernest Valveny Llobet, Dimosthenis Karatzas, "Multimodal Transformer for Comics Text-Cloze", In *IDCAR*, 2024.

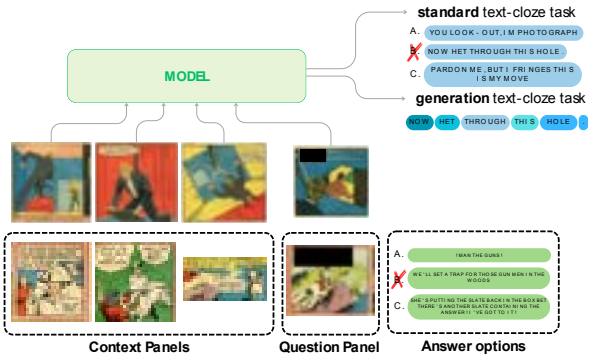


Figure 3: Multimodal model for comics text-close

Emanuele Vivoli received his B.S. degree in Computer Science and Engineering from the University of Florence, Florence, Italy, in 2018. He received his M.S. degree in 2021 in Artificial Intelligence and is currently a Ph.D. Candidate in the Deep Learning and Multimodal Models both at the University of Florence (MICC) and Autonomous University of Barcelona (CVC). His research interests are in the areas of Multimodal learning (Vision and Language), comics and documents understanding, and designing new tasks and datasets in comics.

Computer Vision Techniques for Historical Handwritten Music Recognition

Pau Torras
Advisor: Alicia Fornés
Computer Vision Center & Universitat Autònoma de Barcelona
E-mail: ptorras@cvc.uab.cat

Keywords: computer vision, optical music recognition, handwriting recognition

1 Summary of Previous and Current Work

The main focus of this last year of work has been developing a computer-based notation for music that has certain indispensable properties for the field of Optical Music Recognition (OMR):

- **The representation of an engraved music score must be unique.** If the same score can be represented multiple ways the performance of music recognition models is hard to evaluate fairly.
- **Most widespread approaches for OMR must be supported:** This representation must be applicable regardless of how the notation reconstruction is obtained: Object detection, image-to-sequence, rule-based systems, etc.
- **Music semantics can be inferred from the presentation of the score:** This is important because incorporating certain semantics into the representation poses complex problems. For instance, transposing instruments play notes in a range that is different from the one that is written. Therefore, encoding pitch information directly can be troublesome for end-to-end approaches.

Current music exchange formats do not have these properties. MusicXML or MEI are not permutation-invariant, and they are extremely impractical for use in OMR due to the sheer size of the files they produce. Lilypond and other fast engraving-oriented formats omit relevant details for music notation that are handled by an implicit layout engine. Many other formats are designed to keep only certain aspects of music (pitch information, harmony, etc.).

The solution we devised for this is a tree-based notation system. Its core idea is that music symbols are created composing simple primitives into complex, abstract objects. Therefore, the leave nodes in the tree represent the object primitives present in the score and higher order nodes are its composition. The composition operations are represented using the edges of the tree. Figure 1 shows a detailed example in which a sample score is converted into the format.

As a consequence of these properties, the representation system allows for additional degrees of freedom in terms of scopes of recognition. For instance, it allows the same dataset to be used for recognition at the measure, line or page level independently, because there are no context-spanning objects (provided, of course, some measure-matching algorithm is developed by the user in question). At the same time, it allows to use music transcription or bounding box annotations for the same score.

2 Future Work and Challenges

Once a notation format was established we began working in building a large, comprehensive historical handwritten music dataset. This is ongoing work for which I am providing mostly technical support solving challenges related to music representation and tooling. The bulk of this process is being handled by a team of annotators coordinated by my supervisor and our go-to musicologist Carles Badal. We are also working in an application that allows fast annotation of symbol-level datasets using a handheld tablet device.

In parallel to all of this, the next main challenge is training a moderately competent model using hand-

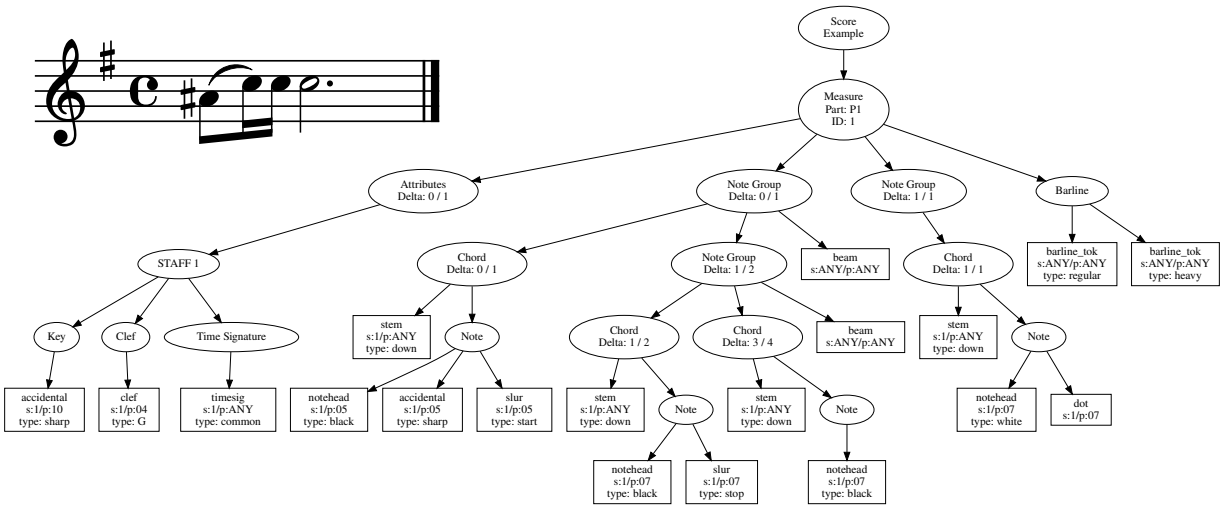


Figure 1: Representation of a score in our music notation format. Each node in the graph represents a primitive in music notation (stem, notehead, rest, etc.) and each edge represents a composition rule towards a higher order object.

written or handwritten-looking data. This is currently in a very early stage for various reasons, the main one being that we do not have a significant amount of properly annotated handwritten scores to train any models from scratch. The roads we are exploring right now involve two steps:

- Pre-training models with large collections of non-annotated scores, which we do have available.
- Generating plausible-looking synthetic scores. We are exploring symbol-level generation techniques for this end in order to have generation for which the ground truth is known.

The main architecture we are exploring for transcription are Transformers because of the wide array of pre-training techniques available to them and their inherent multimodality. There are plenty of models based on Transformers in use for Optical Character Recognition and vision-text tasks such as Visual Question Answering. Moreover, there has been a lot of effort in developing lightweight and efficiently-trained Transformer variants that are suited for the requirements of the task at hand.

The preliminary results we have obtained indicate known truths about Transformers: that they tend to overfit heavily and they are tricky to optimise from scratch. An important aspect for this is finding a good string-based representation of the output such that the output tokens are not very heavily imbalanced in terms of probability of appearance.

Publications

[1] Pau Torras, Sanket Biswas, Alicia Fornés, *"A Unified Representation Framework for the Evaluation of Optical Music Recognition Systems"*. In *IJDAR* (processing) 2024.

Pau Torras received his B.S. degree in Computer Science from Universitat Autònoma de Barcelona, Bellaterra, Spain, in 2021. He received his M.S. degree in Computer Vision in 2022 and is currently a Ph.D. Candidate in the Document Analysis Group at the Computer Vision Center. He received the Best Master's Thesis Award from the Master in Computer Vision in 2022. His research interests are in the areas of computer vision, natural language processing, theoretical computer science and digital humanities. His current work attempts to merge this diverse array of interests into a single tangible line of research on Music Recognition.

Information spotting in sequences and time series: Application to gestures and strokes

Asma BENSALAH
Advisors: Alicia Fornés Bisquerra and Josep Lladós Canet
Computer Vision Center & Universitat Autònoma de Barcelona
E-mail: abensalah@cvc.uab.es

Keywords: Spotting, human movements, deep learning, movement quality, data augmentation for signals, multi-model, neuromuscular diseases.

1 Summary of Previous and Current Work

The application of artificial intelligence (AI) in patient monitoring and disease screening presents various obstacles and constraints. This thesis aims to delve into these challenges and restrictions, with a particular emphasis on patient monitoring, wherein innovative, human-like synthetic generation methods for handwriting have been examined for screening purposes.

Previously, the dataset containing movements of healthy and patient individuals in a constraint and non-constraint scenarios was created. This dataset along with a baseline classification method.

Publications

Bensalah, Asma & Parziale, Antonio & De Gregorio, Giuseppe & Marcelli, Angelo & Fornés, Alicia & Lladós, Josep. (2023). I Can't Believe It's Not Better: In-air Movement for Alzheimer Handwriting Synthetic Generation.

Asma Bensalah, Cristina Carmona-Duarte, Miguel A. Ferrer, Andreas Fischer, Alicia Fornés, Josep Lladós, Eloy Opisso, Réjean Plamondon, Anna Scius-Bertrand and Josep Maria Tormos. 3D Kinematics for Stroke Patient monitoring (submitted).

Asma Bensalah, Alicia Fornés, Cristina Carmona-Duarte and Josep Lladós. Easing Automatic Neurorehabilitation via Classification and Smoothness Analysis. Conference of International Graphonomics Society, IGS 2021. Alicia Fornés, Asma Bensalah, Cristina Carmona-Duarte, Miguel A. Ferrer, Andreas Fischer,Josep Lladós, Eloy Opisso, Réjean Plamondon, Anna Scius-Bertrand and Josep Maria Tormos. The RPM3D project: 3D Kinematics for Remote Patient Monitoring. Conference of International Graphonomics Society, IGS 2021.

Alicia Fornés, Asma Bensalah, Cristina Carmona-Duarte, Miguel A. Ferrer, Andreas Fischer,Josep Lladós, Eloy Opisso, Réjean Plamondon,Anna Scius-Bertrand and Josep Maria Tormos. Exploring the 3D Kinematics for Brain Stroke Rehabilitation

Bensalah, Asma & Chen, Jialuo & Fornés, Alicia & Carmona-Duarte, Cristina & Lladós, Josep & Ferrer, Miguel. (2021). Towards Stroke Patients' Upper-Limb Automatic Motor Assessment Using Smartwatches In Pattern Recognition In ICPR International Workshops and Challenges.

Asma Bensalah received her B.S. degree in Computer Science from Université des Sciences et de la Technologie Houari Boumediene, Algiers, Algeria. She received her M.S. degree in 2019 and is currently a Ph.D. Candidate in the Department of Computational Science at the Universitat Autònoma de Barcelona. Her main interests are gesture recognition, rehabilitation assesment, invasive and non-invasive sensing.

Divide, Ensemble and Conquer: The Last Mile on Unsupervised Domain Adaptation for On-Board Semantic Segmentation

Tao Lian
Advisor: Antonio M. López
Computer Vision Center
E-mail: tlian@cvc.uab.es

Keywords: Unsupervised domain adaptation, semantic segmentation, ensemble

1 Summary of Previous and Current Work

During the past year, my research focused on UDA for semantic segmentation. More specifically, we aim to bridge the last mile, the gap between UDA and supervised learning (SL) with human labelling. Recent UDA methods have progressed significantly, yet they often rely on strategies customized for synthetic single-source datasets (e.g., GTA5), which limits their generalisation to multi-source datasets. Conversely, synthetic multi-source datasets hold promise for advancing the last mile of UDA but remain underutilized in current research.

We propose DEC, a flexible UDA framework for multi-source datasets. DEC follows a divide-and-conquer strategy to train semantic segmentation models and fuse their predictions to generate the semantic segmentation mask. In our first step, we apply the division strategy to train several models on categorized semantic classes, called category models, similarly to how humans semantically label an RGB image using AI tools [1, 2] (first background, then large objects, and finally small ones). The categorization simplifies the number of classes learnt by each category model, expecting a better performance and fast convergence. The next step involves the training of ensemble model, a DeepLabv3+ [3] network, which learns to fuse segmentation masks generated by each category model into one mask. The framework is illustrated in Fig. 1. DEC exhibits compatibility with a wide range of UDA methods, providing consistent improvement. DEC achieves state-of-the-art performance on Cityscapes, BDD100K, and Mapillary Vistas, resulting in mIoU of 78.7, 62.4, and 74.8, respectively. We enhance the existing state-of-the-art reported in [4] by 1.2, 2.1, and 0.6 points, respectively. Our performance is only 2.8 and 4.7 points behind SL for Cityscapes and Mapillary Vistas, surpassing the SL for BDD by 0.9 points. Overall, DEC propels progress in the challenging last mile of UDA through a simple yet effective divide-and-conquer strategy and ensemble method. This work is published on [5].

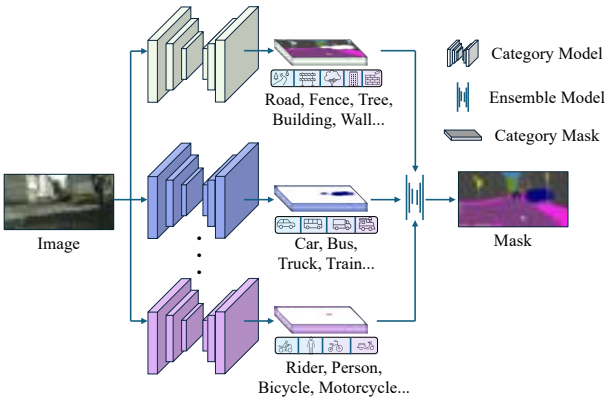


Figure 1: The overview of DEC. It consists of multiple category models and an ensemble model. Category models segment the image into category masks containing distinct classes. Subsequently, the ensemble model fuses these category masks to generate the final mask.

2 Future Work and Challenges

Although the UDA performance on Cityscapes and BDD100K datasets is state-of-the-art, further improvement is needed for Mapillary. Moreover, the performance of domain generalization (DG) still lags significantly behind supervised learning with human labelling. For future work, we plan to apply vision foundation models (VFM) to domain adaptation

and domain generalization for semantic segmentation. This approach leverages the power of VFM, including SAM, DINOV2, and CLIP, to enhance the performance of unsupervised semantic segmentation. The biggest challenge in applying VFM lies in using tokens; for instance, SAM requires specific points or bounding boxes for segmentation. Therefore, determining the appropriate tokens for VFM is crucial for achieving precise segmentation.

Publications

[1] Tao Lian, Jose L. Gómez, Antonio M. López, "Divide, Ensemble and Conquer: The Last Mile on Unsupervised Domain Adaptation for On-Board Semantic Segmentation". In *IEEE Transactions on Intelligent Vehicles* (submitted).

References

[1] D. Acuna, H. Ling, A. Kar, and S. Fidler, "Efficient interactive annotation of segmentation datasets with Polygon-RNN++," in *Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] B. Sambaturu, A. Gupta, C. Jawahar, and C. Arora, "Scribblenet: Efficient interactive annotation of urban city scenes for semantic segmentation," *Pattern Recognition*, vol. 133, p. 109011, 2023.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 834–848, 2017.

[4] L. Hoyer, D. Dai, H. Wang, and L. Van Gool, "Mic: Masked image consistency for context-enhanced domain adaptation," in *Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[5] T. Lian, J. L. Gómez, and A. M. López, "Divide, ensemble and conquer: The last mile on unsupervised domain adaptation for on-board semantic segmentation," 2024.

Tao Lian received his M.S. degree in vehicle engineering from Budapest University of Technology and Economics, Budapest, Hungary. He is currently a Ph.D. candidate advised by full professor Antonio M. López in Computer Vision Center of Autonomous University of Barcelona, Spain. His research interests are autonomous driving, semantic segmentation and unsupervised domain adaptation.

Minimizing Human Labeling in Training Deep Models for Pedestrian Intention Prediction

Muhammad Naveed Riaz
Advisor: Antonio M. López, Maciej Wielgosz
Computer Vision Center & Universitat Autònoma de Barcelona
E-mail: nriaz@cvc.uab.es

Keywords: Vision-based Autonomous Driving, Pedestrian Intention Recognition, Semi-Supervised Learning (SSL), Unsupervised Domain Adaptation (UDA)

1 Summary of Previous and Current Work

Pedestrian intention prediction is a crucial aspect of autonomous driving. It is important to know whether pedestrians are planning to cross in front of the ego vehicle so that it can perform safe and comfortable maneuvers. However, creating fast and accurate models to perform such a prediction from sequential images is challenging. One of the factors that makes it difficult is the lack of datasets with diverse crossing and non-crossing (C/NC) scenarios. To address this problem, we recently introduced PedSynth, a set of synthetic videos that was generated using our CARLA-based framework ARCANE. We trained PedGNN, our new deep model to predict C/NC situations, with PedSynth and real-world datasets (JAAD and PIE) achieving state-of-the-art results [1].

The C/NC ground truth for real-world datasets is manually collected, which can be time-consuming and prone to subjectivity. To overcome this challenge, in this paper, we propose using PedSynth and PedGNN to self-label the real-world training data. The presented results strongly encourage using PedSynth to develop deep models for the C/NC prediction task. In particular, we introduce a procedure inspired by semi-supervised learning (SSL) that allows us to provide real-world videos with C/NC labels automatically. These self-labeled videos and PedSynth are used to train our PedGNN model as part of the SSL procedure. We can see how our self-labeling significantly outperforms human labeling by assessing model performance on real-world testing data. Our study also highlights different inconsistencies found in human-labeled videos. To our knowledge, this paper is the first to assess synth-to-real self-labeling for C/NC prediction. The whole pipeline of our work is depicted in Figure 1.

2 Future Work and Challenges

In our future research, we want to improve the prediction of pedestrian intention (what a pedestrian will do next) by incorporating more information into our model. Currently, our model only uses pedestrian skeletons (the basic stick figure representation of a person's pose) as input. This limits the model's ability to understand the situation fully.

Recent research suggests that using bounding boxes around pedestrians, in addition to skeletons, can improve the accuracy of intention prediction. Bounding boxes provide additional information about the pedestrian's location and size in the image. By combining this information with skeletons using a technique called modality attention, our model can learn to focus on the most relevant parts of each data type for better prediction.

Figure 1: Graphical representation of the presented methodology.

Publications

[1] M. N. Riaz, M. Wielgosz, A. G. Romera and A. M. López, "Synthetic Data Generation Framework, Dataset, and Efficient Deep Model for Pedestrian Intention Prediction". In *IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, 2742-2749, 2023.

Muhammad Naveed Riaz is currently a Ph.D. scholar at the Computer Vision Center (CVC), Autonomous University of Barcelona (UAB), where he works in the field of Autonomous Driving for Pedestrian Safety. He obtained his Master's degree from Shanghai Jiao Tong University (SJTU), China, where his research focus was Facial Expression Recognition. He completed his Bachelor's degree in Computer Science at COMSATS University Islamabad (CUI), Pakistan.

UDA4Inst: Unsupervised Domain Adaptation for Instance Segmentation

Yachan Guo
Advisor: Antonio M. López
Computer Vision Center & UAB
E-mail: yguo@cvc.uab.es

Keywords: Synth-to-Real Unsupervised Domain Adaptation, Instance Segmentation, Vision-based Autonomous Driving

1 Summary of Previous and Current Work

Unsupervised Domain Adaptation (UDA) aims to transfer knowledge learned from a labeled source domain to an unlabeled target domain. While UDA methods for synthetic to real-world domains (synth-to-real) show remarkable performance in tasks such as semantic segmentation and object detection, very few were proposed for instance segmentation in the field of vision-based autonomous driving, and the existing ones are based on a suboptimal baseline, which severely limits the performance. We introduce UDA4Inst [1], a strong baseline of synth-to-real UDA for instance segmentation. UDA4Inst adopts cross-domain bidirectional data mixing at the instance level to effectively utilize data from both source and target domains. Rare-class balancing and category module training are also employed to further improve the performance. It is worth noting that we are the first to demonstrate results on two new synth-to-real instance segmentation benchmarks, with 39.0 mAP on UrbanSyn → Cityscapes and 35.7 mAP on Synscapes → Cityscapes. Our method outperforms the source-only Mask2Former model by +7 mAP and +7.6 mAP, respectively. On SYNTHIA → Cityscapes, our method improves the source-only Mask2Former by +6.7 mAP, achieving state-of-the-art results. The pipeline of our method is shown in Fig. 1.

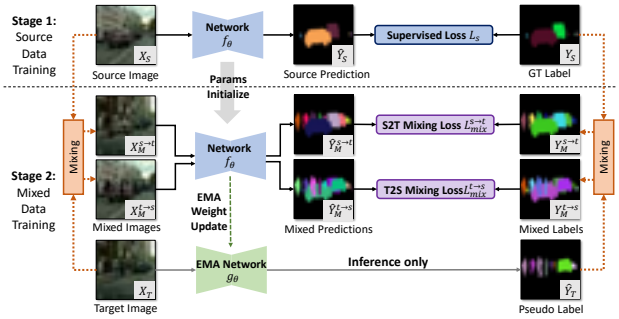


Figure 1: Overview of UDA4Inst framework. In the first stage, the student network is trained with source domain data in a supervised manner. In each iteration of the second stage, we update the EMA teacher g_θ with the weights from the network f_θ to predict pseudo-labels in the target domain. Next, the mixed data are generated with the source and target images, and the mixed labels are generated by the target pseudo-labels and the source ground-truth labels. The network is then trained on the mixed data with S2T and T2S mixing losses.

2 Future Work and Challenges

In our future work, we aim to tackle the significant challenges in accurately segmenting small, distant obstacles, a major issue in unsupervised domain adaptation (UDA) for instance segmentation in autonomous driving. Our research will focus on enhancing feature extraction through multi-scale and image pyramid methods, leveraging large pre-trained models for auxiliary training. Also optimizing instance-level data mixing strategies and designing specialized loss functions for distant obstacles can help in this task. These efforts aim to improve UDA instance segmentation performance and advance autonomous driving technology. These efforts are expected to lead to measurable improve-

ments in segmentation accuracy and robustness, ultimately contributing to safer autonomous driving systems.

Publications

[1] Guo, Y., Xiao, Y., Xue, D., Zurita, J.L.G., Lopez, A.M, UDA4Inst: Unsupervised Domain Adaptation for Instance Segmentation. In *arXiv (2024) https://arxiv.org/abs/2405.09682*.

Guo Yachan received her M.S. degree in 2015 and she has been pursuing her Ph.D. since 2022 at the Autonomous University of Barcelona. Her research interests are in the areas of unsupervised domain adaptation, visual object detection, semantic segmentation, instance segmentation.

Guiding Attention in End-to-End Driving Models

Diego Porres
Advisor: Antonio López Peña
Computer Vision Center & Universitat Autònoma de Barcelona
E-mail: dporres@cvc.uab.es

Keywords: End-to-End Driving, Vision Transformers, Attention

1 Summary of Previous and Current Work

Pure vision-based end-to-end autonomous driving trained via Behavior Cloning (BC) has the advantage of being cheap, as there is no need for costly ground-truth data nor expensive hardware. However, these models still exhibit the shortcomings of end-to-end training, namely in the lack of explainability in the chosen actions.

In the work presented at the IEEE Intelligent Vehicles Symposium 2024 [1], we propose to use *synthetic attention masks* obtained from semantic segmentation images (either already available or from pre-trained models) to optimize the self-attention maps of the Transformer Encoder of CIL++. We do so by matching the distributions of the masks and the self-attention maps. This has several advantages: there is no need to modify the original architecture, the sample efficiency is increased (needing far less amount of data to get the same driving capacity), and the *synthetic attention masks* are not needed during validation.

2 Future Work and Challenges

While the proposed Attention Loss increases the model's interpretability and explainability (Figure 2), we still lack *transparency*. Indeed, not all highlighted objects matter for the current action that the model outputs. Adding this transparency during evaluation is an interesting venue to take. Future work entails learning these *synthetic attention masks* in a self-supervised fashion and using saliency maps from humans that drive inside CARLA.

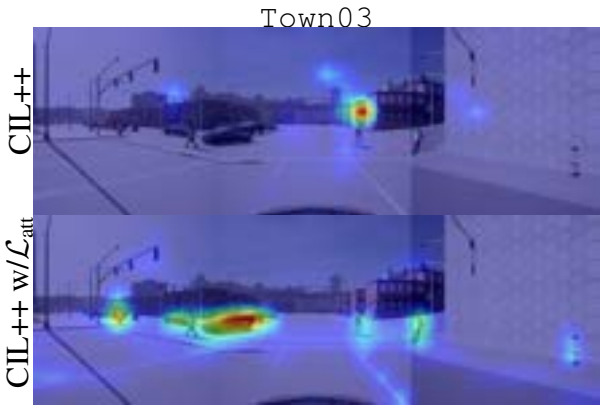


Figure 1: Self-attention maps of the last layer of the Transformer Encoder in the CIL++ architecture. Top: results with self-supervised training. Bottom: By adding the Attention Loss, we can better guide the model in recognizing and segmenting the important objects in the scene.

Publications

[1] Diego Porres, Yi Xiao, Gabriel Villalonga, Alexandre Levy, Antonio M. López. *Guiding Attention in End-to-End Driving Models*. In *IEEE Intelligent Vehicles (IV) Symposium*, 2024.

Diego Porres received his B.S. degree in Physics from Universidad del Valle de Guatemala in 2012, his M.S. in Mathematical Engineering from the Università degli Studi dell'Aquila, Italy, and is currently a Ph.D. Candidate in the Department of Computer Science at the Universitat Autònoma de Barcelona. His research interests are end-to-end autonomous driving, domain adaptation, and unsupervised image generation.

Developing Transparent Artificial Intelligence Mechanisms for Enhanced Decision-Making in Autonomous Vehicles

Alexandre LEVY
Advisors: Antonio Manuel Lopez Peña, Ernest Valveny Llobe
Computer Vision Center & UAB
E-mail: alevy@cvc.uab.cat

Keywords: Autonomous Driving, Artificial Intelligence, Deep Learning, Explainability, Large Language Models

1 Summary of Previous and Current Work

Autonomous Driving (AD) represents a significant leap in automotive technology, where vehicles navigate without human input. AD promises increased safety, efficiency, and accessibility in transportation, yet faces challenges like technological reliability, ethical considerations, and regulatory hurdles. The research that I am focusing on aims to enhance the transparency and reliability of AI systems within autonomous vehicles.

Enhancing transparency in AD involves making AI's decision-making processes more understandable to humans. This includes developing techniques to articulate the rationale behind AI decisions, actions, and predictions. The objectives are twofold: developing explainability models for AI decisions in AD and strengthening AI models' robustness with explainability. These objectives aim to build trust among users, identify and rectify errors or biases, and ultimately contribute to safer and more reliable autonomous vehicles.

2 Future Work and Challenges

When evaluating a newly trained autonomous driving model, it's essential to test it on a variety of predefined routes. While a well-trained model will successfully navigate many of these routes, it will inevitably encounter some failures. To improve the model's performance, it's crucial to retrain it on scenarios similar to those where it failed. To facilitate this process, we have developed a pipeline that generates new training scenarios based on observed failures. Each time the autonomous vehicle encounters a failure, we generate different, yet similar, scenarios to those that caused the failure. This approach helps make our AI driving model more resilient to various challenges. For that, we will simulate everything with CARLA [1] and generate SCENIC [2] Scenarios.

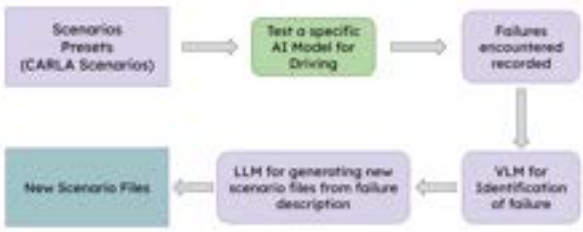


Figure 1: Pipeline of new edge cases generation.

References

[1] A. Dosovitskiy, G. Ros, F. Codevilla, A. L´opez, and V. Koltun, "CARLA: An open urban driving simulator," in Conference on Robot Learning (CoRL), 2017

[2] Fremont, D. J., Dreossi, T., Ghosh, S., Yue, X., Sangiovanni-Vincentelli, A. L., Seshia, S. A. (2019). Scenic: a language for scenario specification and scene generation. In Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation.

Alexandre Levy received his B.S. degree in Microengineering from Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, in 2019. He received his M.S. degree in Robotics in 2022 and is currently a Ph.D. Candidate in the Department of Computer Science at the Universitat Autònoma de Barcelona. His research interests are in the areas of autonomous driving, explainability, large language models and scene generations.

Continually Learning Self-Supervised Representations

Alex Gomez-Villa
Advisors: Joost Van de Weijer, Bartlomiej Twardowski, Andrew Bagdanov
Computer Vision Center & Autonomous university of Barcelona
E-mail: agomezvi@cvc.uab.es

Keywords: Continual learning, self-supervised learning, class-incremental learning

1 Exemplar-free Continual Representation Learning via Learnable Drift Compensation

One of the main limitations of exemplar-free CL methods is that, as the backbone is updated during the training of new tasks, the network experiences semantic feature drift [6]. This feature drift results in catastrophic forgetting and a significant drop in performance. As a consequence, most exemplar-free methods are evaluated in big-start settings (starting from a large first task or from a pretrained network) to start from a strong backbone and thus avoid significant semantic drift of backbone network [6, 2, 8, 7, 9, 4, 3, 5, 1]. In this year, we focus on the exemplar-free small-start setting in which we train a network from scratch, starting with a small task, and continually update the backbone.

We showed that performance degradation in exemplar-free CP accumulation methods is largely due to feature drift. We proposed a simple yet effective prototype correction method called Learnable Drift Compensation (LDC). LDC (Fig.1) learns a projector that maps between the feature spaces of the consecutive tasks using only the data available at each task along with current and previous model. The trained projector is used at the end of each task to correct the stored old prototype positions. We demonstrate the efficacy of LDC in correcting the drift of any moving backbone by using it on top of already established exemplar-free CL strategies. Extensive experiments on CIFAR-100, Tiny-ImageNet, and ImageNet100 show that LDC outperforms previous drift correction methods in supervised and semi-supervised settings.

2 Future Work and Challenges

- Review, study, and propose metrics to evaluate color perception inside visual-language models
- Review, study, and propose methods to mitigate semantic drift in generative models.

References

[1] Liu, Xialei, Chenshen Wu, Mikel Menta, Luis Herranz, Bogdan Raducanu, Andrew D Bagdanov, Shangling Jui, and Joost van de Weijer. Generative feature replay for class-incremental learning. In *Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[2] Dipam Goswami, Yuyang Liu, Bartłomiej Twardowski, and Joost van de Weijer. Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *International Conference on Computer Vision*, 2019.

[4] Mark D McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton van den Hengel. Ranpac: Random projections and pretrained models for continual learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[5] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Conference on Computer Vision and Pattern Recognition*, 2019.

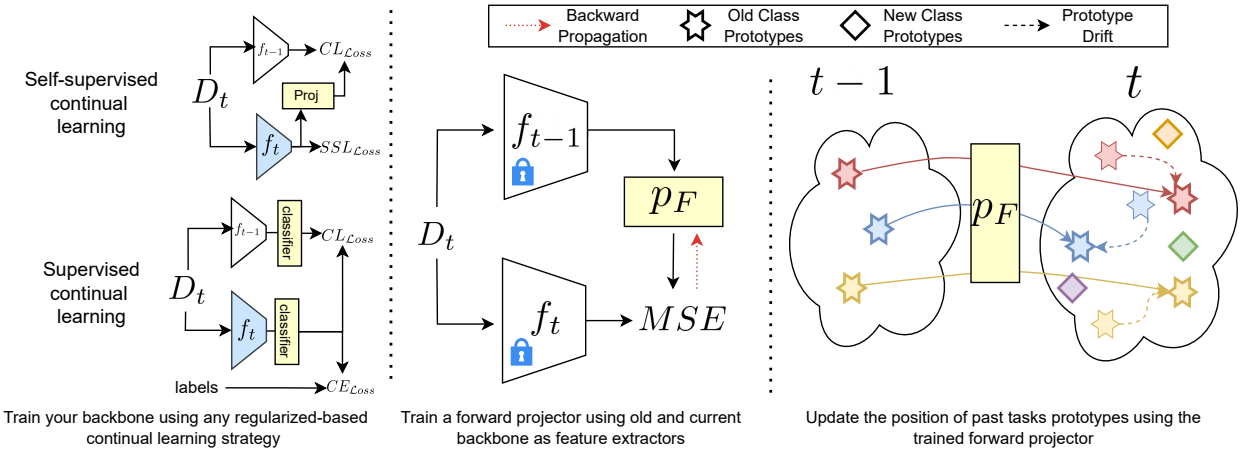


Figure 1: **Learnable Drift Compensation**. We train the model on the current task data using regularization-based, self-supervised or supervised continual representation learning methods. After training the new feature extractor f_t, we learn a forward projector p_F by minimizing the mean squared error between the projected features from f_{t-1} and the features from f_t. We use the learned projector to compensate for the drift of the old class prototypes in the new feature space.

[6] Lu Yu, Twardowski, Bartlomiej, Liu, Xialei, Luis Herranz, Kai Wang, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6982–6991, 2020.

[7] Fei Zhu, Zhen Cheng, Xu-yao Zhang, and Cheng-lin Liu. Class-incremental learning via dual augmentation. *Advances in Neural Information Processing Systems*, 34:14306–14318, 2021.

[8] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5880, 2021.

[9] Kai Zhu, Wei Zhai, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Self-sustaining representation expansion for non-exemplar class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9296–9305, 2022.

Alex Gomez-Villa received her B.S. degree in Electrical Engineering from University of Antioquia, Medellin, Colombia, in 2016. She received his M.S. degree in 2018 and is currently a Ph.D. Candidate in the Department of Computer Science at the Autonomous University of Barcelona. Her research interests are in the areas of computer vision, continual learning, and human visual perception.

Exemplar-Free Continual Learning

Dipam Goswami
Advisor: Joost van de Weijer, Bartłomiej Twardowski
Computer

Keywords: Class-Incremental L
lanobis Distance, Adversarial Attack

1 Summary of Previous rent Work

I have been working actively in the tinual learning, particularly the exem where we are not allowed to store da tasks and learn only using the curren tinual Learning (CL) is a machine le that focuses on adapting models to while retaining previous information

We exploit the heterogeneity o tions in exemplar-free continual lea per [1]. Existing methods in contin ically use the euclidean metric for feature embeddings. Since the classe trained in continual learning settings, the euclidean metric might not be the optimal one. The objective is to study the anisotropic Mahalanobis distance for classification in continual learning settings. This can be applicable to many-shot, few-shot or even in domain-incremental settings. We propose in our paper, a non-linear Mahalanobis-distance based classifier which exploits the feature covariances of the classes and achieves state-of-the-art results across many-shot, few-shot CL benchmarks and CL with pre-trained ViTs.

We also explore semantic drift estimation methods for Continual learning. Current drift estimation methods works well with good representation from the first task model trained on 50% of the dataset. However, this is not so effective in more practical settings when the first task has less number of classes. The objective is to study and propose a robust drift compensation method for class-incremental learning without using exemplars and training from scratch instead of a big first task. We propose in our recent paper [2], how adversarial samples can be generated using targeted adversarial attacks on new task samples which can serve as pseudo-exemplars of old classes. The proposed method is depicted in Fig. 1. We use these adversarial samples for prototype drift compensation and achieve state-of-the-art performance on equal-split small-start settings on multiple datasets.

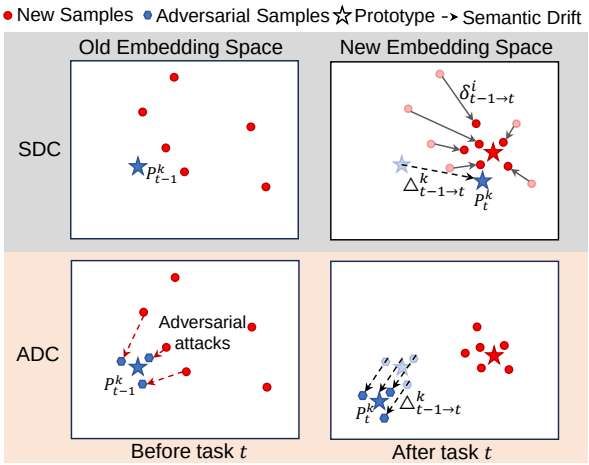


Figure 1: Illustration of our method - Adversarial Drift Compensation (ADC) and existing method - SDC. In SDC, the drift $\Delta^k_{t-1\rightarrow t}$ is estimated as the average of drift of all new task samples after training on a new task. Instead, we propose to move the new task features close to the old prototype P^k_{t-1} of class k by perturbing the new images using targeted adversarial attacks. The drift of the adversarial samples from old to new feature space is used to resurrect all old prototypes.

2 Future Work and Challenges

We plan to continue our research on continual learning. We are also exploring the inclusion of language models in our research. Considering both LLM and

•Remove t
•Change 'a
'**perturbed**
•Increase t
the clumsi

VLM continual learning.

Publications

[1] Dipam Goswami, Yuyang Liu, Bartłomiej Twardowski, and Joost van de Weijer, Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[2] Dipam Goswami, Albin Soutif-Cormerais, Yuyang Liu, Sandesh Kamath, Bartłomiej Twardowski, and Joost van de Weijer, Resurrecting old classes with new data for exemplar-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Dipam Goswami received his B.E. degree in Computer Science and M.Sc. degree in Mathematics from BITS Pilani, India, in 2022. He is currently a Ph.D. Candidate at the Computer Vision Center and the Department of Computer Science at the Universitat Autonoma de Barcelona. His research interests are in the areas of continual learning with a focus on data- and label-efficient approaches and federated learning.

Generative models in Cross-Domain Image Processing

Patricia L. Suárez
Advisor: Angel D. Sappa
Computer Vision Center & Universitat Autònoma de Barcelona
E-mail: 1508312@uab.cat

Keywords: generative models, cross-domain imaging, multispectral imaging

1 Summary of Previous and Current Work

During my studies in the PhD program, I have conducted research on techniques for dehazing, colorizing, and synthesizing images of different spectra, as well as improving image resolution. Recent research has proposed transformer-based techniques to achieve super-resolution or dehaze and/or synthesize images. These approaches improve results by integrating data from multiple spectral bands, overcoming the limitations of single-band (visible spectrum) imaging. This multispectral integration significantly improves classical tasks such as object detection and recognition, edge detection, and classification. Working with the translation of information from different spectra presents challenging problems that require innovative solutions, during the last academic year my research mainly focused on generating depth images and synthetic thermal images using cross-domain generative networks, see Fig. 1 and Fig. 2 respectively. The central idea is to use depth maps as prior information to improve the quality of the synthesized thermal images. Continuing with my research work, I have explored the fusion of depth images with the edges of real thermal images to create a combined image that is used as input to the synthesis model. According to the results of the experiments, this fusion image is the one that has given the best results in the synthesis of both image characteristics and temperatures. The proposed approach has been evaluated in different datasets (e.g., FLIR ADAS V2, M3FD and our Thermal Stereo data set). With this, it has been able to verify the effectiveness of generative models in cross-domain image processing.

2 Future Work and Challenges

Current work includes evaluating and comparing different depth map models to be used as information that conditions the generation of thermal-like images. At the moment we plan to explore the implementation of a loss that aims to capture and penalize structural differences between the generated images and the real images. The central idea is that this loss function serves to evaluate the similarity between the generated image and the real image not only in terms of pixels but also in terms of spatial characteristics, which is a challenge considering that there is no direct correlation between the depth of the images versus the temperatures. However, with this loss, the aim is to improve the pseudothermal image synthesis model, based on the fact that it improves the capture of image characteristics. Also, I will start writing my doctoral thesis.

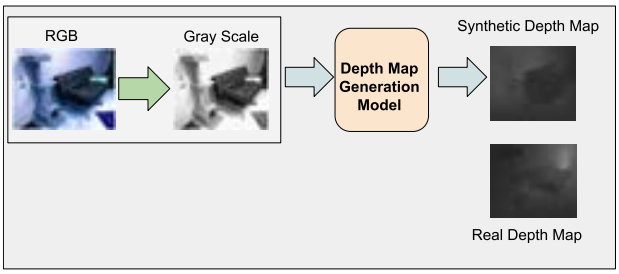


Figure 1: Depth Map Synthetizing Architecture.

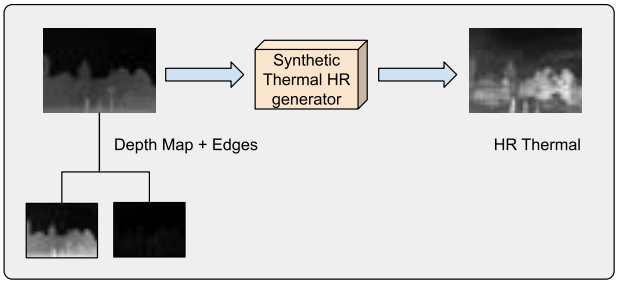


Figure 2: Thermal Synthetizing Architecture.

Publications

[1] P. L. Suárez and A. Sappa, "Depth-Conditioned Thermal-like Image Generation," in *IEEE 14th International Conference on Pattern Recognition Systems (ICPRS)*, London, England, 2024.

[2] Suárez, P.L., Carpio, D., Sappa, A. (2023). A Deep Learning Based Approach for Synthesizing Realistic Depth Maps. In: *Foresti, G.L., Fusiello, A., Hancock, E. (eds) Image Analysis and Processing. ICIAP 2023. Lecture Notes in Computer Science*, vol 14234. Springer, Cham

[3] P. L. Suárez, D. Carpio and A. Sappa, "Boosting Guided Super-Resolution Performance with Synthesized Images," 2023, In *17th International Conference on Signal-Image Technology - Internet-Based Systems (SITIS)*, Bangkok, Thailand, 2023, pp. 189-195

[4] P. L. Suárez, D. Carpio and A. Sappa, "Depth Map Estimation from a Single 2D Image", In *2023 17th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS)*, Bangkok, Thailand, 2023, pp. 347-353.

[5] Suárez, P. L., and Sappa, A. D. "A Generative Model for Guided Thermal Image Super-Resolution", In *4th International Conference on Innovations in Computational Intelligence and Computer Vision (ICICV)*, Jaipur, India.

[6] Suárez, P. L., Carpio, D., and Sappa, A. D. (2024) "Enhancement of guided thermal image super-resolution approaches", *Neurocomputing*, 573, 127197

Patricia L. Suárez graduated as a Computer Systems Engineer in 2002, and obtained a Master's Degree in Management Information Systems in 2008 from ESPOL. She is a Ph.D. candidate in Computer Vision Center at the Universitat Autònoma de Barcelona. She received the Best Paper Award at CVPR-PBVS workshop in 2017. She has 12 years of experience in the professional field in the different computer science sectors. Since 2006, has been actively involved as a professor at different universities in Ecuador, sharing her knowledge and experience with her students. Additionally, since 2015 she has been a researcher in the field of processing and representation of multispectral images at the CIDIS research laboratory at ESPOL, Guayaquil, Ecuador. She has been a research team member on national and international projects.

ColorPeel: Color Prompt Learning with Diffusion Models via Color and Shape Disentanglement

Muhammad Atif Butt
Advisor: Joost Van De Weijer
Computer Vision Center & Universitat Autonoma De Barcelona
E-mail: mabutt@cvc.uab.es

Keywords: Diffusion Models, Color Prompt Learning, Generative AI

1 Summary of Previous and Current Work

Text-to-Image (T2I) generation has made significant advancements with the advent of diffusion models. Previously, we focused on prompt based image editing using text-to-image models, where we aim to change the visual appearance of a target object (or objects) in the image without any unintentional modifications. We proposed Dynamic Prompt Learning [1] to force cross-attention maps to focus on correct noun words in the text prompt while updating the dynamic tokens for nouns in the textual input with the proposed leakage repairment losses to achieve fine-grained image editing over particular objects. We use several automatic metrics to quantitatively compare our method against existing baselines, and further affirm the results using a user study.

This year, we propose to learn specific color prompts tailored to user-selected colors. Naive color prompt learning approaches tend to result in color-shape entanglement. To overcome this, we generate several basic geometric objects in the target color, allowing for color and shape disentanglement during the color prompt learning, as depicted in Fig. 1. Our method, denoted as ColorPeel [2], successfully assists the T2I models to *peel off* the novel color prompts from these colored shapes. In the experiments, we demonstrate the efficacy of ColorPeel in achieving precise color generation with T2I models. Our findings represent a significant step towards improving precision and versatility of T2I models, offering new opportunities for creative applications and design tasks.

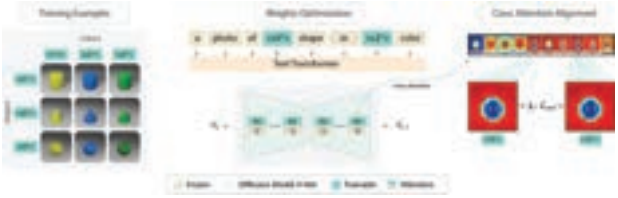


Figure 1: Illustration of our method ColorPeel. Instance images along with the templates are generated, given the user-provided RGB or color coordinates. Next, we introduce new modifier tokens, i.e., $< s >$ and $< c >$ which correspond to shapes and colors to ensure the disentanglement of shape from color. To improve learning, we introduce *cross attention alignment* to enforce the color and shape cross-attentions.

2 Future Work and Challenges

In future research, we aim to explore learning a color grid of tokens, allowing users to interpolate the color combinations, to avoid training for new colors. Since color is a continuous concept, adding tokens to the CLIP dictionary alone cannot provide a comprehensive solution. This means that representing all possible colors through a single training process is still a challenging task, and we plan to address this in future work.

Publications

[1] Kai Wang, Fei Yang, Shiqi Yang, Muhammad Atif Butt, Joost van de Weijer, Dynamic Prompt Learning: Addressing Cross-Attention Leakage for Text-Based Image Editing. In *NeurIPS*, 2023.

[2] Muhammad Atif Butt, Kai Wang, Javier Vazquez Corral, Joost van de Weijer, ColorPeel: Color Prompt Learning with Diffusion Models via

Color and Shape Disentanglement. In *ECCV*, 2024.

Muhammad Atif Butt is currently a Ph.D. candidate at Computer Vision Centre, The Autonomous University of Barcelona. He received his master's degree of Computer Science in 2022 from National University of Science and Technology (NUST), Islamabad, Pakistan, and bachelor's degree of Computer Science in 2018 from Mirpur University of Science and Technology (MUST), Mirpur, Azad Kashmir, Pakistan. Before starting Ph.D. at CVC - UAB, he worked as Research Associate at National Centre of Robotics and Automation, Pakistan, and Information Technology University, Lahore, Pakistan. His research interests are in the areas of computer vision and intelligent information systems; particularly, applied artificial intelligence (AI), and generative modelling.

Sketching Imagination: Explorations of Generative Models for Visual Creativity

Ayan Banerjee
Advisor: Josep Lladós Canet
Computer Vision Center & Universitat Autònoma de Barcelona
E-mail: abanerjee@cvc.uab.es

Keywords: Sketch Generation, Diffusion Model, Graph Neural Networks

1 Summary of Previous and Current Work

Generating vector graphics from text prompts is a challenging vision task, requiring diverse yet realistic depictions of the seen as well as unseen entities. However, existing research has been mostly limited to the generation of single objects, rather than comprehensive scenes comprising multiple elements. In response, this thesis introduces SVGCraft (see Figure 1), a novel end-to-end framework for creating vector graphics depicting entire scenes from textual descriptions. Utilizing a pre-trained LLM for layout generation from text prompts, this framework introduces a technique for producing masked latent in specified bounding boxes for accurate object placement. It introduces a fusion mechanism for integrating attention maps and employs a diffusion U-Net for coherent composition, speeding up the drawing process. The resulting SVG is optimized using a pretrained encoder and LPIPS loss with opacity modulation to maximize similarity. Additionally, this thesis explores the potential of primitive shapes in facilitating canvas completion in constrained environments.

2 Future Work and Challenges

One problem that we are currently facing is that there are no text-sketch pair datasets. So it isn't easy to train a network in a supervised way. So we are using the pretrained combination of the diffusion models to solve this challenge.

In future work, we are planning to use a sketch prompt instead of a text prompt to generate 3D objects and comics. As sketch provides more control over text prompts towards generation it will be an interesting direction to explore further.

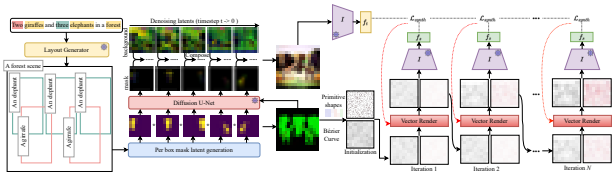


Figure 1: Graphical representation of SVGCraft.

Publications

[1] Banerjee A, Biswas S, Lladós J, Pal U. SemiDocSeg: harnessing semi-supervised learning for document layout analysis. International Journal on Document Analysis and Recognition (IJDAR). 2024 Jun 4:1-8.

[2] Banerjee A, Biswas S, Lladós J, Pal U. GraphKD: Exploring Knowledge Distillation Towards Document Object Detection with Structured Graph Creation. ICDAR 2024.

Ayan Banerjee received his B.S. degree in Computer Science & Engineering from JGEC, India in 2021. He received his M.S. degree in 2024 and is currently a Ph.D. Candidate in the Department of Computer Science at the Universitat Autònoma de Barcelona. His research interests are in the areas of generative models, and ad-hoc networks with emphasis on mathematical modeling and performance analysis.

Synthesizing Aesthetic Diversity: Creating a Balanced Dataset of Visual Aesthetics Using GANs

Arslan Javed
Advisor: Carlos Alejandro Parraga, Bogdan Raducanu
Computer Vision Center & UAB
E-mail: ajaved@cvc.uab.cat

Keywords: Aesthetic Perception, Visual aesthetics, Generative Adversarial Networks (GANs), Synthetic Dataset.

1 Summary of Previous and Current Work

The field of visual aesthetics mainly emphasizes the study of beautiful and visually pleasing images, often neglecting the aspect of ugliness in the existing datasets. However, understanding the perception of ugliness is crucial for a comprehensive grasp of aesthetic perception as a whole. The insufficient data on ugliness in visual aesthetics hinders advancements in comprehending the full spectrum of human aesthetic perception and restricts the development of accurate algorithms and models that can analyze and evaluate the visual appeal of various images, including those deemed aesthetically unpleasant or unattractive.

To address this problem, we have made significant achievements in expanding and refining our dataset of aesthetically diverse natural images, particularly focusing on the aspect of ugliness. Leveraging Generative Adversarial Networks (GANs), Specifically the StyleGAN2-ADA[2] model, we have generated a comprehensive and balanced dataset categorized into five different aesthetic classes, from ugly to beautiful. The fidelity and realism of generated images were validated using different metrics such as Freshet Inception Distance (FID), t-distributed Stochastic Neighbor Embedding(t-SNE), and Confusion Matrix as shown in Figure 1. Additionally, we performed detailed analyses of low-level image features to check the properties of both real and generated images. we employed techniques such as outlier removal, cosine similarity[1], and color similarity-based image filtering to maintain the diversity and integrity of the dataset.

1.1 Identify particular images using Beta Distribution and study characteristics:

Furthermore, In our dataset each image was evaluated by 100 individuals, leading to some cases where votes for aesthetic ratings were highly polarized, categorizing the same image as both 'ugly' and 'beautiful'. To address this, we employed a beta distribution to model the voting results. This approach helped us effectively identify images with polarized ratings from the rest of the dataset. This was done to study the characteristics of these images, aiming to uncover the underlying factors that cause viewers to perceive the same image as both ugly and beautiful, providing insights into the complex nature of aesthetic perception.

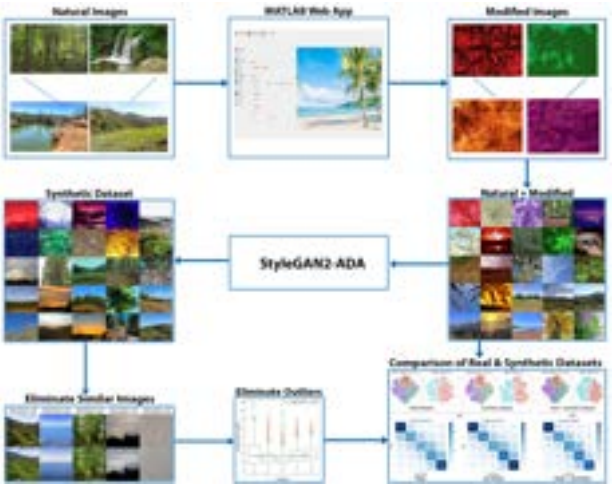


Figure 1: Graphical representation of the presented methodology.

The comprehensive dataset created through this research serves as a valuable resource for future studies in visual aesthetics. By addressing the often-neglected aspect of ugliness, this work contributes to a more holistic understanding of aesthetic percep-

tion.

Moreover, we are planning to collect the images that are uncomfortable for humans and compare the properties of ugly and uncomfortable images, aiming to understand the distinctions and overlaps between these two aspects of visual perception. This comparison will provide a deeper insight into the cognitive and emotional responses enriching our understanding of human aesthetic perception. This approach ensures that our findings are robust, applicable, and valuable for future advancements in computer vision and aesthetic analysis.

2 Future Work and Challenges

Moving forward, there are several research directions and challenges that we anticipate in our current state of research:

a) Evaluation of the particular images: We are conducting a thorough analysis of characteristics for the images that are categorized as equally 'ugly' and 'beautiful' by 100 individuals. This included both low and high level features of the images to uncover the underlying factors that cause half of the observers to perceive an image as ugly and the other half as beautiful.

b) Comparative analysis of Ugly and Uncomfortable Images: In the next phase of our research, we plan to explore the properties of ugly and uncomfortable images to understand the distinctions and overlaps between these two aspects of visual perception. Uncomfortable images are those that deviate from statistical properties of natural scenes, particularly the scale-invariant distribution of Fourier energy. These images tend to concentrate visual energy at certain spatial frequencies, particularly around three cycles per degree, causing strain and discomfort to the visual system[3]. Moreover, recent research provides valuable insights into how specific visual features influence psychological responses, which is crucial for understanding why certain images may be perceived as uncomfortable [4]. This comparison of ugly and uncomfortable images will involve a detailed analysis of the cognitive and emotional responses these images provoke. We will investigate various low-level image features such as color, contrast, texture, and composition to identify patterns and unique characteristics of each category. Our goal is to uncover the underlying factors that contribute to the perception of discomfort and ugliness, providing deeper insights into human aesthetics perception.

By addressing these research directions and challenges, our aim is to advance our understanding of aesthetic perception and image analysis. Through the exploration of various aspects of aesthetic perception, including the perception of ugliness and other dimensions such as beauty, uncomfortable, and visual appeal, we seek to gain a comprehensive understanding of how humans perceive and evaluate visual stimuli.

Arslan Javed received his B.S. degree in Computer Science from COMSATS University Pakistan, where he was awarded the silver medalist distinction. He then pursued his M.S. degree at the University of Electronic Science and Technology China (UESTC) from 2017 to June 2019. In January 2023, he began his Ph.D. studies at the Computer Vision Center (CVC) UAB where he focuses on topics related to computer vision, image processing, and perception. His research aims to advance our understanding of human perception and develop innovative algorithms and models for computer vision applications.

References

[1] Mominul Islam, Hasib Zunair, and Nabeel Mohammed. Cossif: Cosine similarity-based image filtering to overcome low inter-class variation in synthetic medical image datasets. *arXiv preprint arXiv:2307.13842*, 2023.

[2] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

[3] Olivier Penacchio and Arnold J. Wilkins. Visual discomfort and the spatial distribution of fourier energy. *Vision Research*, 108:1–7, 2015.

[4] Meinald Thielsch and Gerrit Hirschfeld. Facets of website content. *Human-Computer Interaction*, 34:1–49, 03 2018.

Transferring Generative Models

Héctor Laria Mantecón
Advisor: Joost van de Weijer, Bogdan Raducanu
Computer Vision Center & UAB
E-mail: hector.laria@cvc.uab.es

Keywords: Continual Learning, Neural Networks, Deep Learning, Computer Vision, Transfer Learning

1 Summary of Previous and Current Work

In our research methodology, we focused on efficient low-rank adaptations applied to continual learning. Low-rank adaptations are a popular alternative to finetuning of large pretrained models. Similar to regular continual finetuning, we explore better continual approaches for these adapters. We found that indirect regularization of these adapters, yet more resource efficient, cannot surpass existing direct regularization strategies for continual concept learning.

Fundamental knowledge forgetting study of low-rank adaptations of foundational image generation models. Application of low-rank adapters helps modifying models to new or particular behaviors. However, there has not been any exhaustive study on how incorporation of these adapters affect existing knowledge in the model. We are following this idea, in its application to text-to-image generative models.

2 Future Work and Challenges

Future research will delve into exploring and harnessing the representational power of Diffusion Models, which have demonstrated remarkable capabilities in content synthesis. However, significant challenges lie in understanding how these representations are laid out in the generative model and how they can be edited, corrected or replaced effectively. Addressing these issues will be a key focus in overcoming the current challenges of efficient utilization of Diffusion Models.

Figure 1: Representation of how continual concept learning should happen when using low-rank adaptations. Reality is harder, since destructive interference among concepts may arbitrarily occur.

Publications

[1] NeRF-Diffusion for 3D-Consistent Face Generation and Editing. Héctor Laria, Kai Wang, Joost van der Weijer, Bogdan Raducanu. *VISAPP 24*.

Héctor Laria received his B.S. degree in Madrid, Spain in 2017. He obtained his M.S. degree from Aalto University, Helsinki, Finland in 2019 and is currently pursuing a Ph.D. in the LAMP group at Universitat Autònoma de Barcelona. His research focuses on Deep Learning, Computer Vision and Generative Models with a particular emphasis on Generative Adversarial Networks, Diffusion Models and Transfer Learning.

Adversarial Attacks on Latent Diffusion Models

Xide Xu
Advisor: Bogdan Raducanu
Computer Vision Center
E-mail: xide@cvc.uab.es

Keywords: Image Generation, Diffusion Model, Adversarial Attack, Privacy Protection

1 Summary of Previous and Current Work

The primary focus of this research work has been to explore methods that enhance privacy protection in text-to-image diffusion models. This involves investigating techniques to ensure data confidentiality and security, minimizing the risk of information leakage, and implementing robust privacy-preserving mechanisms. The motivation for this research stems from the powerful capabilities of new token learning diffusion models, such as Custom Diffusion and DreamBooth. This characteristic has raised significant privacy concerns if used in an unauthorized way.

Custom Diffusion[4] and DreamBooth[3] represent significant advancements in the personalization of text-to-image generation models. Custom Diffusion allows for fine-tuning the model to generate images that closely align with specific user requirements. By leveraging a few reference examples, it can learn new tokens and integrate them into the diffusion process, enabling highly personalized outputs. DreamBooth further enhances this personalization by fine-tuning the model with minimal examples, resulting in exceptionally high-quality and tailored image generation.

However, the strengths of these models also pose substantial risks to personal privacy. Their ability to generate realistic images from limited data can lead to unauthorized use of personal information and artwork. To address these privacy concerns, MIST[1] introduces adversarial perturbations by using the AdvDM algorithm, which optimizes latent variables from the reverse process of diffusion models to prevent imitation of artistic styles. Anti-DreamBooth[2] uses a surrogate DreamBooth model trained on a small subset of clean data to guide adversarial attacks, finding optimal noise to perturb target images and prevent accurate replication by other DreamBooth models. These works all use the Projected Gradient Descent (PGD)[5] to generate adversarial examples. PGD is a well-established adversarial attack method that perturbs input data to maximize the model's prediction error within a specified bound. Figure 1 and Figure 2 show examples of MIST and Anti-Dreambooth.

Figure 1: Examples of MIST attacking method. The first row is the perturbed images with different noise levels, and the second is the images generated by Dreambooth trained using the perturbed images

One of the main drawbacks of methods such as Anti-Dreambooth and MIST is that they attack the whole model, targeting all the parameters, which is totally inefficient. Recent research[6] has found that cross-attention layers are particularly critical in these models, as they undergo significant parameter changes during training and play a pivotal role in the model's ability to generate coherent and contextually accurate images. By targeting these layers, the aim is to disrupt the mapping between text and image during the fine-tuning of customized diffusion models. The underlying idea is to freeze the model parameters except W_K and W_V (which handle the text input

Figure 2: Examples of Anti-Dreambooth attacking method. The first row is the perturbed images with different noise levels, and the second is the images generated by Dreambooth trained using the perturbed images.

features) and only update them to facilitate the learning of the mapping between text input and image input by the model. After this disruption, the model is capable of recognizing the content of the image, but it lacks the ability to categorize or classify it into specific groups or types.

Motivated by the previous findings, we have recently started to look into the fabrics of cross-attention layers, and fully identify and exploit their vulnerabilities in order to increase the model's robustness against adversarial attacks.

2 Future Work and Challenges

The future work will involve a detailed analysis of cross-attention maps before and after PGD attacks to determine how these perturbations influence the model's attention mechanism and overall performance. These insights will help us to increase the privacy and security of personalized data.

Additionally, we will explore the transferability of privacy-preserving techniques across various diffusion models, assessing and adapting adversarial methods to ensure effectiveness. Extending these techniques to multi-concept diffusion models and image data, like those in Custom Diffusion, is crucial for maintaining privacy in complex scenarios.

Developing robust defense mechanisms against adversarial attacks will also be a key focus. This includes exploring modifications to the attention mechanism and implementing more resilient training protocols to address identified vulnerabilities.

By addressing these challenges and exploring these future research lines, we aim to contribute significantly to the development of secure and robust text-to-image generation models that protect user privacy and intellectual property in the rapidly evolving field of generative AI.

References

[1] Liang, Chumeng. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *ICML*, 2023.

[2] Van Le. Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. In *ICCV*, 2023.

[3] Ruiz. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023.

[4] Kumari. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023.

[5] Aleksander Madry. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*, 2018.

[6] Jingyao Xu. Perturbing Attention Gives You More Bang for the Buck: Subtle Imaging Perturbations That Efficiently Fool Customized Diffusion Models. In *CVPR*, 2024.

Xide Xu received his B.S. degree in Mechanical and Automation Engineering from Huaqiao University, Xiamen, China, in 2019. He received his M.S. degree in 2023 at Xiamen University, Xiamen, China. He is currently a Ph.D. Candidate (First year) in the Computer Vision Center of Universitat Autonoma de Barcelona His research interests are in the areas of generative model, and image attacking.