

Navigating the Data Science Landscape: Personal Insights and Tips

Christian E. Westermann

University of Southern Denmark

November 17, 2023

About Me

- Name: Christian E. Westermann.
- Interests: Sports (tennis) and everything Computer Science/Tech-related.
- PhD-student in Machine Learning and Data Science.

Beginning of Academic Journey



- Quickly ditched Biomedicine.
- Regression Analysis (Statistics): Sparked my interest.
- Thesis in predicting outcomes of football matches using Neural Networks.

Masters and PhD



- RA in Computer Vision.
- Consultancy in Finance.
- Internship in Finance doing NLP.
- Taught Natural Language Processing, Statistical Learning and Data Driven Decision Making
- Pivoted to a PhD in ML and Data Science (ETA summer 2024).

Interests in Data Science and Machine Learning

- Experience showed the extreme versatility within the field → Foundations!
- Introduced to Computer Science part of the field.
- Deep Learning.
- High Performance Programming: C/C++, Assembly, CUDA C.
- Bayesian Statistics.
- I strive to be a generalist. (!)

TableParser: Automatic Segmentation of Historical Documents

A photograph of a yellowed, rectangular historical document titled "Swedish Gradesheet". The document contains handwritten student names and grades in columns. A large red mark "X" is visible at the bottom left.

Figure: Swedish Gradesheet

A photograph of a US Census form from 1950. The form is a grid with various sections for demographic information. It includes handwritten responses such as "Florida" for state and "Black" for race. There are also several redacted sections indicated by black boxes.

Figure: US Census

- Why not just use the Vision tools online?

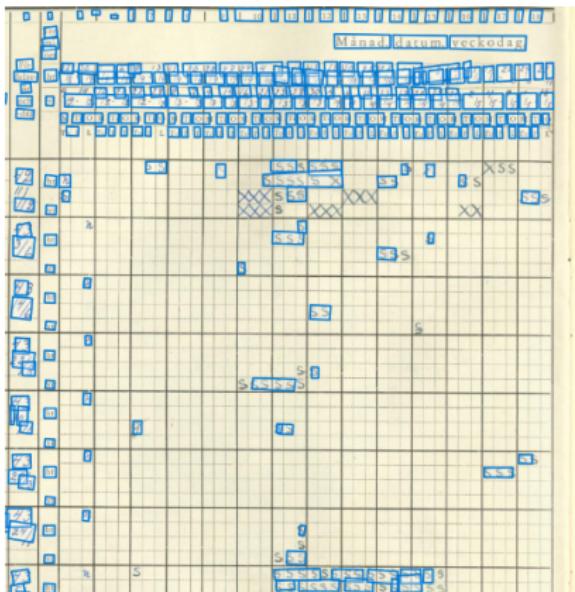


Figure: Swedish Gradesheet (Vision Studio)

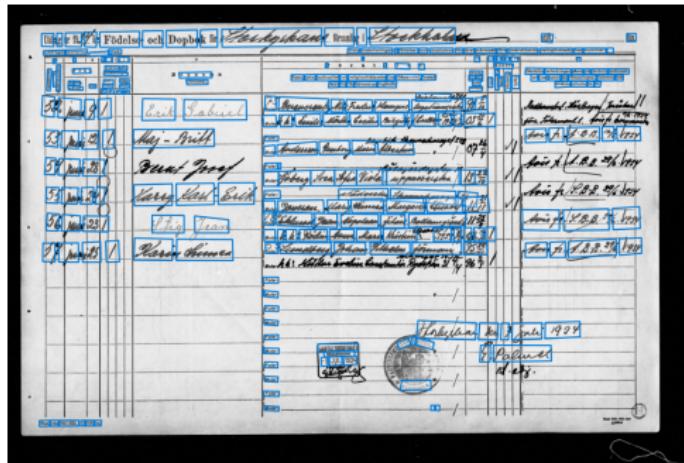


Figure: US Census (Vision Studio)

- They might even look like this

Utdrag ur 100 års Födelse- och Döpbok för Storkyrkoförsamlingen i Stockholm kontrakt Peal

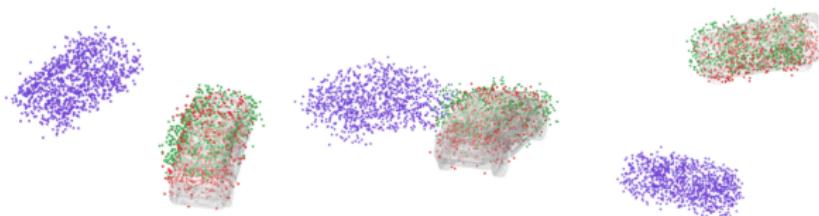
Födelse- och Döpboken för det närmast före ovanstående redogörelse står med hörande årsnummer.

Utdrag omfattar 15 blad.

År och månad	dag	FÖDSEL	Biskop till förfaslägnet	Döpt namn (fornamn)	FÖRTÄGELSE	Namn, yrke, nationitet och religiöskänslan (om faktumade)	Föddes år, dag och måned	Modren	Särskilda anteckningar, särst. om medreg nellerstet (om annan än förfaslägnet), inkomna och afstånd attester m. s.			
									År	m	År	m
1	Jan. 1		Sven Edward (äxta)	Sven Edward	Fader: Okänd,	Krigssoldat. 11	-					
					Moder: Ahlström	Sia Margareta Sophie	87 2/2					
2	4	6	Knut Alexander (äxta)	Knut Alexander	Fader: Okänd.	Böschungs. 5	-					
					Moder: Adolfsson Hilda Louise Cecilia	77 4/3						
3	8	1	Axel Waldemar	Axel Waldemar	Fader: Forsman, Jonas Waldemar Svenn.	75 4/3						
					Moder: Gisselsson, Maria Vilhelmina	77 4/3						
4	8	1	Göta Waldemar	Göta Waldemar	Fader: Petersson, Richard Waldemar, Klocke	79 7/3						
					Moder: Kahlund, Nanny Maria Praggy	78 3/3						
5	10	8	Knut Herbert (äxta)	Knut Herbert	Fader: Okänd.	-						
					Moder: Simonsen, Elena Maria, Kartmarg.	84 6/3						
6	12	6	Greta Emmy Linnea (äxta)	Greta Emmy Linnea	Fader: Okänd.	Brunungs. 1.	-					
					Moder: Petersson, Emma Charlotta	87 1/3						
7	13	1	Anna Visa	Anna Visa	Fader: Dahl, Robert Salomon, Detektörskon.	72 2/3						
					Moder: Genberg, Dr. Fredrika, Skomak. 3	84 4/3						
8	15	8	Augusta Vilhelmina Selma (äxta)	Augusta Vilhelmina Selma (äxta)	Fader: Okänd.	Gartmanns. 21	1888					
					Moder: Höglunds, Augusta Karolina Josephina	80 1/3						
9	16	6	Gustaf Axel (äxta)	Gustaf Axel (äxta)	Fader: Okänd.	Böschungs. 33.	-					
					Moder: Wermack, Matilda Carolina	91 1/2						
10	24	1	Sven Pa	Sven Pa	Fader: Hartzell, John Paul Knudsen.	79 2/3						
					Moder: Bertha Hartzell, Wessinborg	90 3/3						

TableParser: Automatic Segmentation of Historical Documents

- Precise transcription of large and dense tabular documents is highly dependent on the quality of the segmentation of the tables.
- Propose a precise, fast and robust table identification and segmentation pipeline of tabular documents:
 - Semantic segmentation: Unet and SegFormer deep learning architectures for identifying table lines (on pixel level)
 - Point registration aligning target/template and source: FilterReg (likelihood based) and PCRNet (deep learning based)



TableParser: Overview

B

1) Patient Navn Fødselsdato og død dato Dato fødtes med andre tilhørende Uph. gik. Endnu ikke		Richard Adelmaas Hansen name	
2) Født:		Fødested: <u>Se læge</u> birth_date birth_place	
3) Billed og Håndskrift:		Billedet og håndskriften skal overlægnes med andre tilhørende og overlægnes med andre tilhørende om denne dokumentation ikke bliver accepteret.	
4) Brug af Mælktidens Omtale:		Kontrol af Mælktidens Omtale: <input checked="" type="checkbox"/> Ja <input type="checkbox"/> Nej	
5) Døds:		Kontrol af Mælktidens Omtale: <input checked="" type="checkbox"/> Ja <input type="checkbox"/> Nej	
6) Dødsstedsbetegnelse:		Kontrol af Mælktidens Omtale: <input checked="" type="checkbox"/> Ja <input type="checkbox"/> Nej	
7) Dødsstænger og de samtidige symptomer ved døden:		Kontrol af Mælktidens Omtale: <input checked="" type="checkbox"/> Ja <input type="checkbox"/> Nej	
8) Symptomene Variget:		Kontrol af Mælktidens Omtale: <input checked="" type="checkbox"/> Ja <input type="checkbox"/> Nej	
9) Dødsdag: Denne regnskab med dødsstænger og symptomer er ikke nødvendig.		Kontrol af Mælktidens Omtale: <input checked="" type="checkbox"/> Ja <input type="checkbox"/> Nej	
10) Underskrift Lægepræst:		Underskrift Lægepræst: <u>Richard Adelmaas Hansen</u> Lægepræst: <u>Richard Adelmaas Hansen</u> Præstekode: <u>1234567890</u>	
11) Underskrift patient:		<u>Richard Adelmaas Hansen</u> Patient: <u>Richard Adelmaas Hansen</u> Præstekode: <u>1234567890</u>	
12) Underskrift over at denne dokumentation ikke er udarbejdet med andre tilhørende:		<u>Richard Adelmaas Hansen</u> Dokumentation: <u>Richard Adelmaas Hansen</u> Præstekode: <u>1234567890</u>	
13) Underskrift over at denne dokumentation ikke er udarbejdet med andre tilhørende:		<u>Richard Adelmaas Hansen</u> Dokumentation: <u>Richard Adelmaas Hansen</u> Præstekode: <u>1234567890</u>	
14) Underskrift over at denne dokumentation ikke er udarbejdet med andre tilhørende:		<u>Richard Adelmaas Hansen</u> Dokumentation: <u>Richard Adelmaas Hansen</u> Præstekode: <u>1234567890</u>	
15) Underskrift over at denne dokumentation ikke er udarbejdet med andre tilhørende:		<u>Richard Adelmaas Hansen</u> Dokumentation: <u>Richard Adelmaas Hansen</u> Præstekode: <u>1234567890</u>	
Se Anvisninger på Bagside.			

Target

B.		Deceased patient at all times.	
<input type="checkbox"/> Father <input checked="" type="checkbox"/> Mother <input type="checkbox"/> Son <input type="checkbox"/> Daughter <input type="checkbox"/> Brother <input type="checkbox"/> Sister		name _____	
<input type="checkbox"/> Yes <input checked="" type="checkbox"/> No <input type="checkbox"/> Don't know		Relationship to deceased patient _____	
<input type="checkbox"/> Father <input checked="" type="checkbox"/> Mother <input type="checkbox"/> Son <input type="checkbox"/> Daughter <input type="checkbox"/> Brother <input type="checkbox"/> Sister		Date of birth _____ Year _____	
<input type="checkbox"/> Yes <input checked="" type="checkbox"/> No <input type="checkbox"/> Don't know		Place of birth _____	
<input type="checkbox"/> Yes <input checked="" type="checkbox"/> No <input type="checkbox"/> Don't know		Occupation _____	
<input type="checkbox"/> Yes <input checked="" type="checkbox"/> No <input type="checkbox"/> Don't know		Residence _____	
<input type="checkbox"/> Yes <input checked="" type="checkbox"/> No <input type="checkbox"/> Don't know		Death date _____	
<input type="checkbox"/> Yes <input checked="" type="checkbox"/> No <input type="checkbox"/> Don't know		Occupation _____	
<input type="checkbox"/> Yes <input checked="" type="checkbox"/> No <input type="checkbox"/> Don't know		Death cause _____	
<input type="checkbox"/> Yes <input checked="" type="checkbox"/> No <input type="checkbox"/> Don't know		Disease duration _____	
<input type="checkbox"/> Yes <input checked="" type="checkbox"/> No <input type="checkbox"/> Don't know		Death signs _____	
Remarks _____ If another language other than Swedish is used _____ Signature _____ Date _____			
Signature _____ Date _____ Relationship to deceased patient _____ Signature _____ Date _____			

Source

B.		Dedsattest udstdt af en Læge.	
(Denne datsattest skal ikke underskriftes til disse under 1. Art, eftersom i Tidsskriftet om Salmede øste der et oplysningsbrev fra - Jon D. Big. 1912)		Pris: Kr. 10-0	
<p><u>1) Fødsels-Navn (Se alle 25 Aarsdage)</u> <u>2) Fødselsdato (Se fødselsdato med højre Pigenavn, hvis Internationale dato)</u></p>		<p><u>Narren Birthe f. Petersen</u> <u>C. 1880</u></p>	
<p><u>3) Pædi</u></p>		<p>Fødselsdato <u>17/11/1880</u> birth_date birth-place <u>Birket ved Hagle</u> <u>near Roskilde</u> </p>	
<p><u>4) Billed og Hvervnavn (Egen, Efternavn, Fornavn)</u> <u>5) Hoved- og side-tilhørighed (Se hvilke slægtskab, hvilke slægter, hvilke slægtfædre, hvilke slægtmoders, hvilke slægtfædre, hvilke slægtmoders)</u></p>		<p>Occupation <u>Købmand (Handelsmand)</u> <u>Residence</u> <u>Roskilde</u> <u>Roskilde</u> </p>	
<p><u>6) Død</u></p>		<p>Death <u>1912</u> death_date age <u>75</u> <u>75</u> </p>	
<p><u>7) Dødsdiagnose (Se nærmere vedhæftning)</u></p>		<p>Causes <u>Alte</u> death_cause <u>helse</u> disease_duration </p>	
<p><u>8) Begraphens Vedhæftning</u></p>		<p>Death signs <u>Udskriften</u> death </p>	
<p><u>9) Begravelse (Se nærmere vedhæftning)</u> <u>10) Begravelsesplads (Se nærmere vedhæftning)</u></p>		<p>Understyrte Leges navn d. <u>29/11/1912</u> vedt. Ligt d. <u>29/11/1912</u> <u>Janne Birthe f. Petersen</u> da denne sagsættes under et strækkende Deckeblad <u>Lyngby</u> d. <u>29/11/1912</u> <u>A. Borstel</u> Legens hoved i Arbejderforeningens Dommer </p>	
<p>Idebetragt: <u>Arbejdslæge</u> Bogen Kildemand d. <u>29/11/1912</u></p>		<p>Præsenteret: <u>Jens & Bertha B.S.</u></p>	
<p>Justeret: <u>Jens & Bertha B.S.</u></p>		<p>Præsenteret: <u>Jens & Bertha B.S.</u></p>	

Aligned

TableParser: Overview

Foto Nr. 84

B. Dødsattest udstedt af en Læge.

(Dette Bladet må ikke brygges til Barn under 1 År, ejderer i Tidslinje af Selskab eller Døst ved vigtig Hændelse — (b. Lov af 4. Maj 1870))

1) Patientens Navn (først Navn, Søn af, Middelnavn, Efternavn)	Robert Adelma Hansen	
2) Fødested og -dato	Fødested 30. Junii	Fødested Taastrup
3) Dødsstedsnavn	i Byen København Christiania var Lei af ægtske hænses' børne bokse	
4) DødsstedsAdresse (Husnr./Gade, By, Postnummer, Dato)	København Finsborgsgade 79 K. Landeb. Nr. 599 Lei av ægtske hænses' børne bokse	
5) Dødsdato	Dødsdag 29. Aug.	Ålder 6 år
6) Dødsårsagen (Hvor Døden skete først, dvs. først tegnene eller først døden)	København (Husnr./Gade, By, Postnummer, Dato) — Landeb. Nr. 599 Lei av ægtske hænses' børne bokse	
7) Dødsattestens navn og underskrift	Præsidenten over Rigshospitalets Mortalitetsafdeling Robert Adelma Hansen	
8) Dødsattestens Vejleder	6 Aar	
9) Dødsattestens underskrift (Ligesom den der underskrevet, men med lidt længere)	Præsidenten over Rigshospitalets Mortalitetsafdeling Robert Adelma Hansen	
10) Underskrift af Legat d. Robert Adelma Hansen	Præsidenten over Rigshospitalets Mortalitetsafdeling Robert Adelma Hansen	
11) Underskrift af Legat d. Robert Adelma Hansen	Præsidenten over Rigshospitalets Mortalitetsafdeling Robert Adelma Hansen	
12) Se Anmærkninger på Baguden.		

Se Anmærkninger på Baguden.

Robert Adelma Hansen

name

birth_date birth_place

i Byen København Christiania var
Lei af ægtske hænses' børne bokse

occupation

København Finsborgsgade 79 K. Landeb. Nr. 599
Lei av ægtske hænses' børne bokse

residence

disease_duration age

Lei av ægtske hænses' børne bokse

age

København (Husnr./Gade, By, Postnummer, Dato) — Landeb. Nr. 599
Lei av ægtske hænses' børne bokse

occupation

Lei av ægtske hænses' børne bokse

death_cause

Lei av ægtske hænses' børne bokse

disease_duration

Lei av ægtske hænses' børne bokse

death_signs

TableParser: Overview

- Estimate reference document structure Y .
- Specify table cells of interest, Z .
- Estimate target document structure, X .
- Probabilistic point-to-point-set registration (href)
 - Leads to a motion parameter vector $\theta = (\theta_{rot}, \theta_{trans})$, which aligns $X(\theta)$ and Y .
 - Extract cells from image, i.e., Z from $X(\theta)$.
 - Maximize the following loglikelihood function.

$$L = \sum_{i=1}^M \log\left(\sum_{j=1}^{N+1} P(y_j) p(x_i(\theta)|y_j)\right)$$

What Are These Point Clouds and How Do We Get Them?

- We need point clouds that are approximately the same across documents and contain spatial information → table outlines.
- Semantic Segmentation models to go from image to cloud.
 - UNet (Ronneberger et al. 2015)
 - SegFormer (Xie et al. 2021)
 - Both with advantages and disadvantages.

Training Semantic Segmentation Models

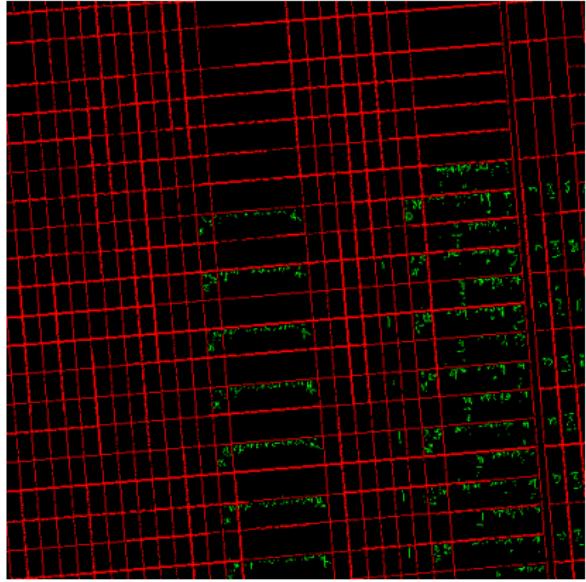
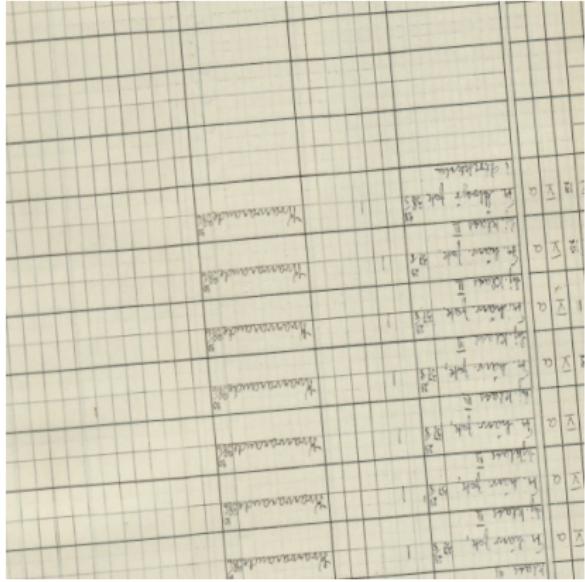
d. 19%. Lärjungarnas omsättning i läraravdelningen under redovisningsåret.

Anmerkungen:

Observera anvisningarna i slutet av boken!

d. % 19~~35~~. Lärjungarnas omsättning i läraravdelningen under redovisningsåret.

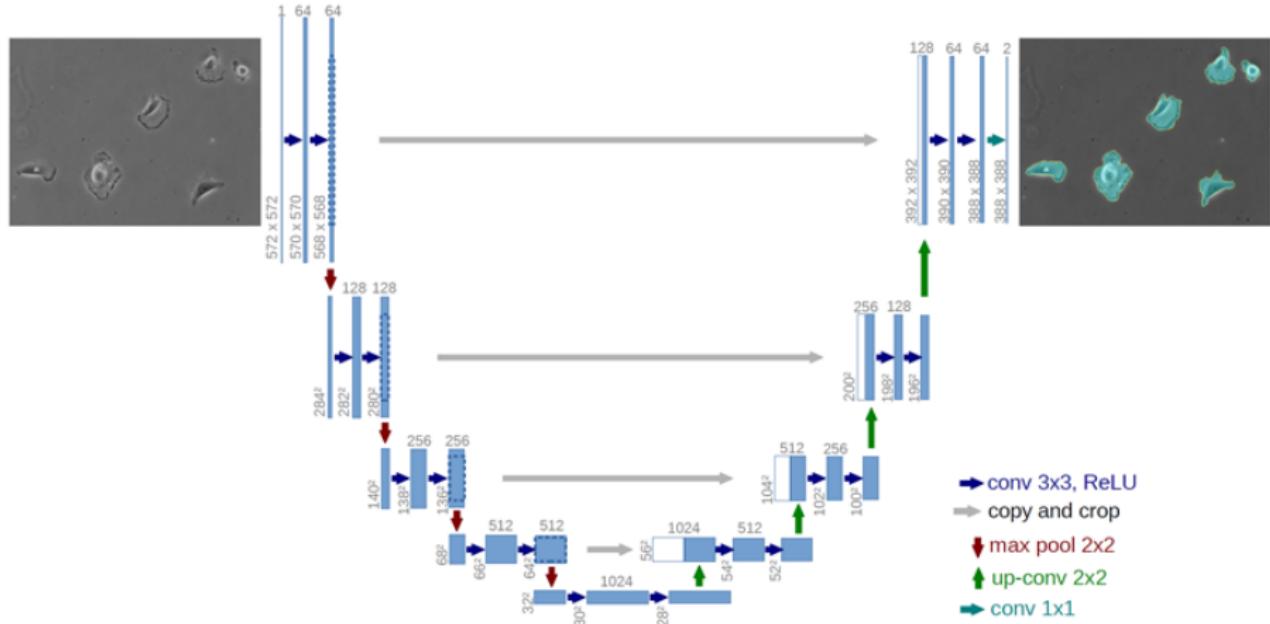
Training Semantic Segmentation Models: Augmentations



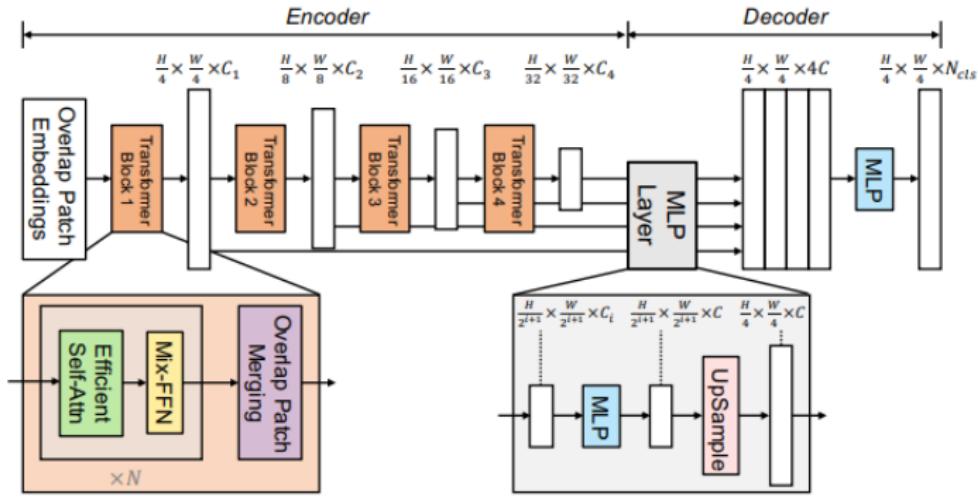
Training Semantic Segmentation Models: Computational Issues

- Documents easily reach approximately 4000×6000 pixels in size.
- That's 24 million variables.
- Vision models usually work in the ballpark of 224×224 to 512×512 .
- Solution: Crop out 512×512 images randomly, which has several benefits
 - Less variability but more examples → better generalizability.
 - Combining random augmentations, different full scale documents and random crops (with random scaling) → an enormous data set.

Training Semantic Segmentation Models: UNet



Training Semantic Segmentation Models: SegFormer



Training Semantic Segmentation Models: OOS Prediction

Barnets namn (Först tillnamn, därefter minst ett förnamn)	Ölund Andreas Andreas, Ingemar 431
Matrikelbladets inskrivningsnr	
Dagur och tennar, då undervisat	
Utdräff för läsåret	
minstens spårdom	
meddelande	
annan meddelande	
Följande visades senare: För konflikter och läsförståelse: Rödhet = A, Blå modersmålet, Inga goda svar = B, Goda svar = C. Lila färger = D. Lila färger = E. För matematik: Rödhet = A, Grönhet = B, Blå modersmålet, Inga goda svar = C, Goda svar = D. För geografi: Myskhet grot = A, Grot = B, Moders grot = C. För ordning: Myskhet grot = A, Grot = B, Moders grot = C.	

Vitsord											
Ter	mi	nner	Uppfrände	Ordnings	Kristendomskunskap	Tal- och läs-	svynningar	Moders-	mat-	Rätskrivning	Geografi
ht	A	AB	Ba	Ba	Ba	AB	AB	Ba	AB	AB	AB
vt	A	A	NB	NB	NB	AB	AB	AB	AB	AB	AB
ht	A	N	Ba	Ba	Ba	Ba	Ba	AB	AB	AB	AB
vt	A	A	Ba	AB	Ba	Ba	Ba	Ba	Ba	Ba	Ba

1	2
Matrikelbladets inskrivningsnr	Barnets namn (Först tillnamn, därefter minst ett förnamn)
2959	Ölund Andreas Andreas, Ingemar 431

Applying the Models

Sliding the Window

- Slide the window left to right and top to bottom.
- At every frame, do semantic segmentation.
- Compare point clouds using Chamfer Distance

$$d_{CD}(S_1, S_2) = \frac{1}{|S_1|} \sum_{x \in S_1} \min_{y \in S_2} \|x - y\|_2 + \frac{1}{|S_2|} \sum_{y \in S_2} \min_{x \in S_1} \|y - x\|_2.$$

Applying Cells Directly vs. After Transformation

Chamfer Distance: 0.0884

Chamfer Distance: 0.0071

A New Data Scientist (Personal Insights and Tips)

Disclaimer: These are personal suggestions and not going one route or the other does not mean you're worse off!

- Data Science and ML content and development is ever increasing.
- Learn how to learn!
- Fundamentals, fundamentals, fundamentals.
 - Within Programming
 - Within Math (Stats, Linear Algebra, Calculus, Probability)
- It's a marathon, not a sprint.
- Think: What could you see yourself working with in the future?

Mathematics (Personal Insights and Tips)

- Foundations
 - Statistics
 - Probability
 - Linear Algebra
 - Calculus
- Figure out whether you benefit most from reading or listening (or both).
- Check out 3Blue1Brown on Youtube for some amazing series.
- Apply what you learn!

Data Analysis (Personal Insights and Tips)

- What areas/domains are you passionate about?
 - Medicine (computer vision, signal processing, ...)
 - Environment.
 - News.
- Search for data sets.
- Do tiny analyses.
- Kaggle competitions.

Deep Learning (Personal Insights and Tips)

- Consider learning about Autograd
 - Implement a toy version yourself (just scalar)
- Learn partial derivation.
 - Also in the case of broadcasting!
 - Dead neurons.
 - Inefficient training.
- Why is weight initialization important, why do we need it?
- Art of stacking well-behaved differentiable blocks.
- Check out Andrej Karpathy's youtube.
- The annotated transformer:
<https://nlp.seas.harvard.edu/2018/04/03/attention.html>
- Try and implement foundational papers.

Programming (Personal Insights and Tips)

- Don't get paralyzed by the "correct" way to code.
- Learn a language with an interpreter (Python, R, ...).
- Learn a language with a compiler (C/C++, Zig, Rust, Java, ...)
- Projects, projects, projects, don't get stuck in tutorial hell.
 - What interests you? Hobbies, scientific/non-scientific fields, etc.
 - See if you can combine it with a programming project.
 - If you're learning a new language: Program something you've done before.
- Why could low-level programming be important to me?
 - Where others hit a wall, you will not.
 - Less reliance on other departments.
 - Be able to create fast and *good* programs.
- If this doesn't interest you, no worries!

On ChatGPT and LLMs (Personal Insights and Tips)

- The GPS of critical thinking.
- Don't take everything it says at face value.
- A templating machine and enhanced google.
- It's not deterministic.

General Tips

- Start small, it's okay to get overwhelmed.
- Find your area of interest (will require a lot of experimenting)
- Find a mentor, many people gladly lend a hand because they likely got help too at some point.
- Track your time, where does it go?

Resources

ML Math:

- Mathematics for Machine Learning. (Deisenroth et al.)

Linear Algebra:

- Linear Algebra Done Right (Sheldon Axler)
- Linear Algebra and Learning from Data (Gilbert Strang)

Statistics / ML

- Introduction to Statistical Learning: <https://www.statlearning.com/> (free)
- All of Statistics
- Probabilistic Machine Learning: An Introduction. (free)
<https://probml.github.io/pml-book/>

Deep Learning (Field moves so fast)

- Dive Into Deep Learning. <http://d2l.ai/index.html> (free)

Q&A