Data Quality In The Cultural Heritage Sector: From An Image Processing Perspective

Tan Lu

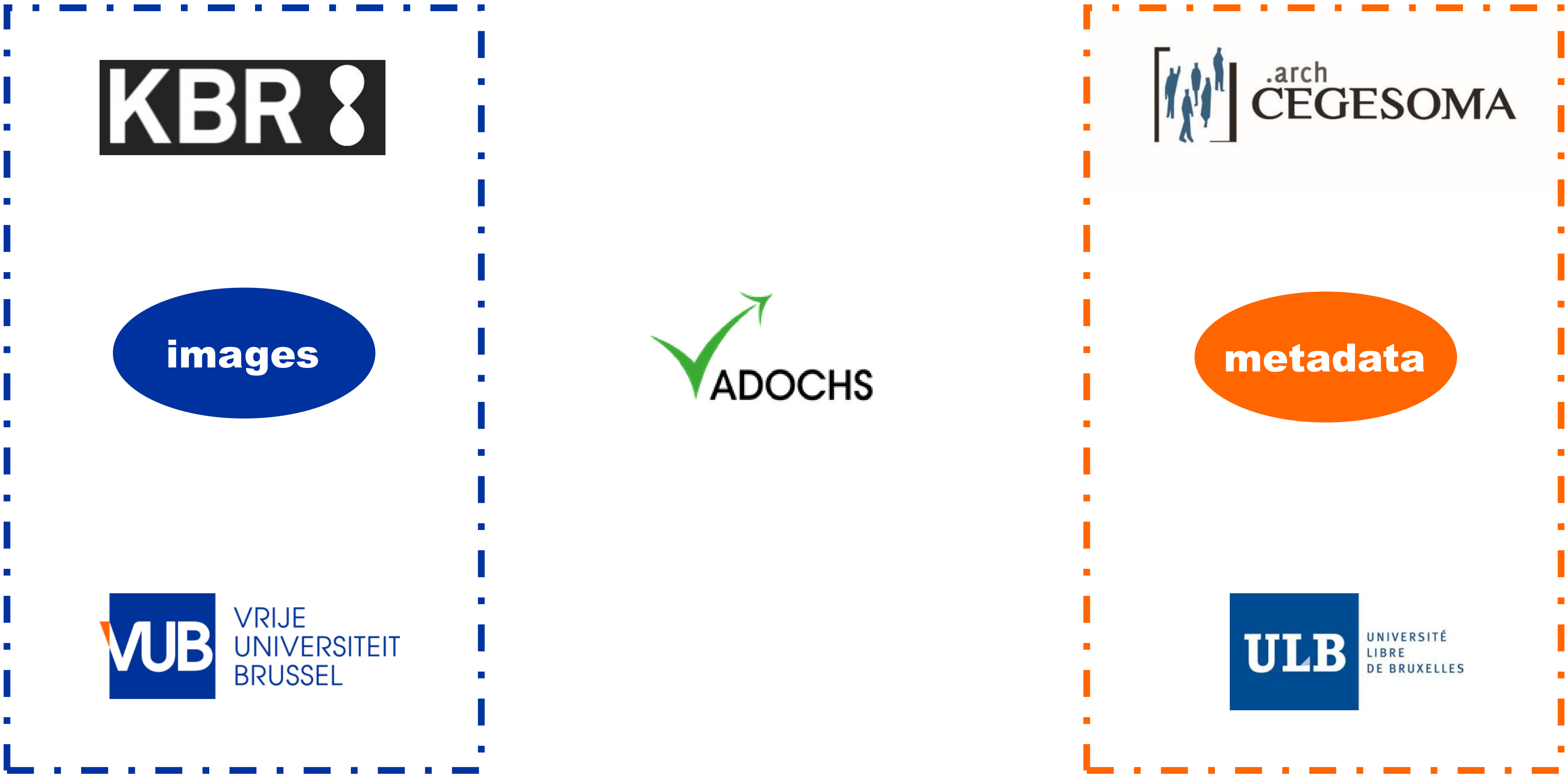
Department of Mathematics and Data Science

Vrije Universiteit Brussel (VUB)

Department of Digitization

Royal Library of Belgium (KBR)

The ADOCHS Project

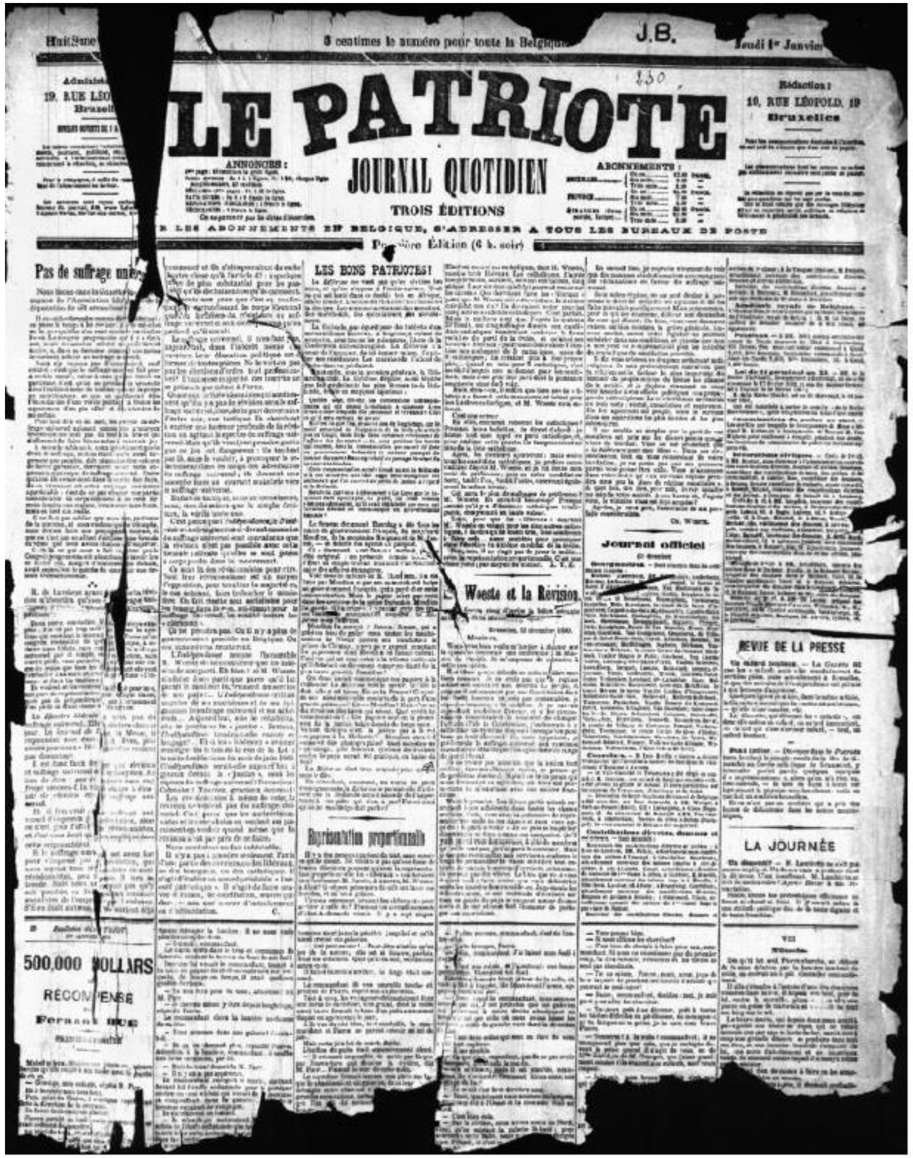


images

metadata

KBR
.arch CEGESOMA
ADOCHS
VUB VRIJE UNIVERSITEIT BRUSSEL
ULB UNIVERSITÉ LIBRE DE BRUXELLES

Images – Digitization in the Cultural Heritage Sector

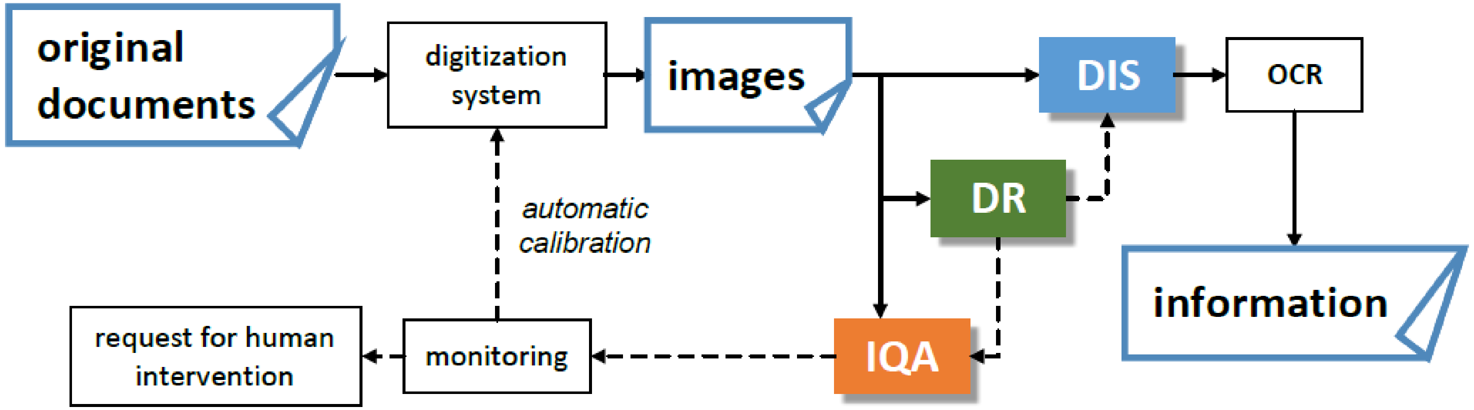


LE PATRIOTE

JOURNAL QUOTIDIEN

TROIS ÉDITIONS

- ***Digitization in the Cultural Heritage Sector***
  - ❑ ***Google Books***: over 40 million books
  - ❑ ***Europeana Newspapers***: aggregating 18 million historic newspaper pages and converting 10 million newspaper pages to full text
  - ❑ ***Royal Library of Belgium (KBR)***: 4500 medieval codices and about one million prints and drawings
- ***Challenges and Opportunities***
  - ❑ ***Image Quality Assessment (IQA)***
  - ❑ ***Image Understanding***
    - Document Image Segmentation (DIS)
    - Damage Recognition (DR)



original documents

digitization system

images

DIS

OCR

DR

information

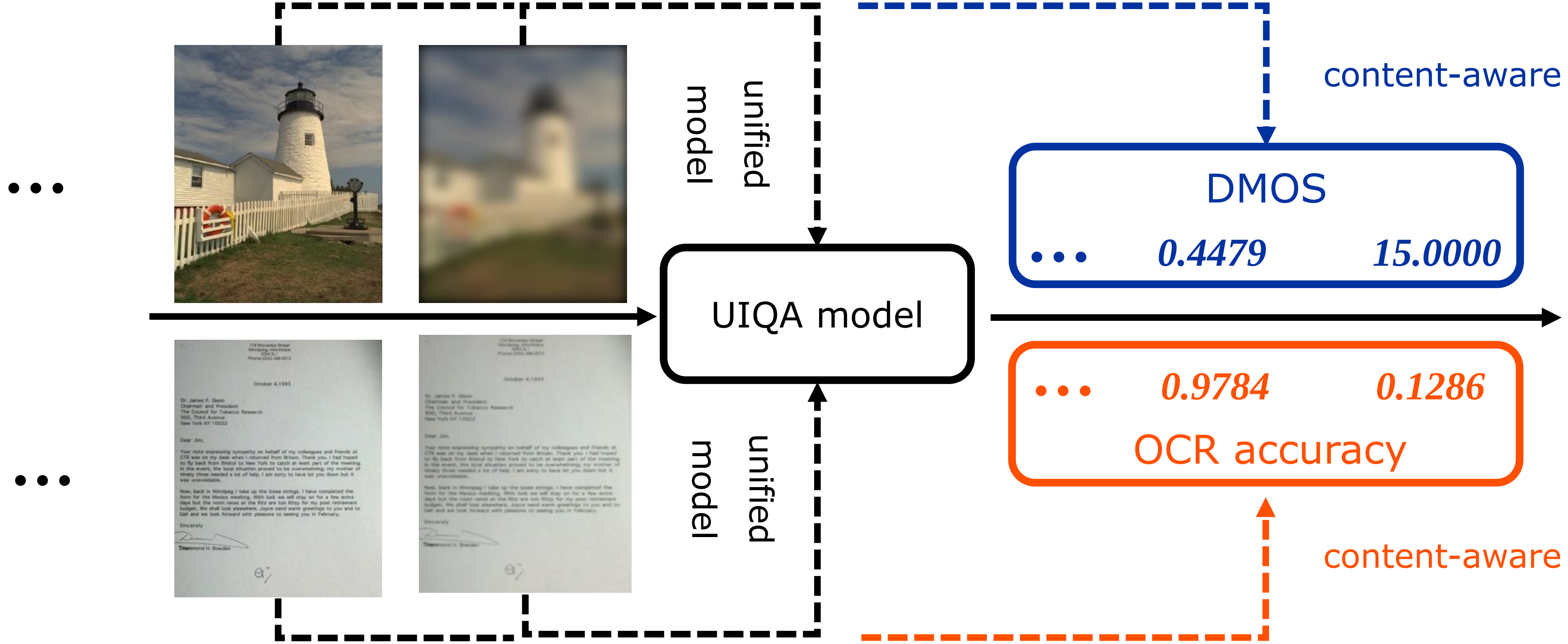
automatic calibration

request for human intervention

monitoring

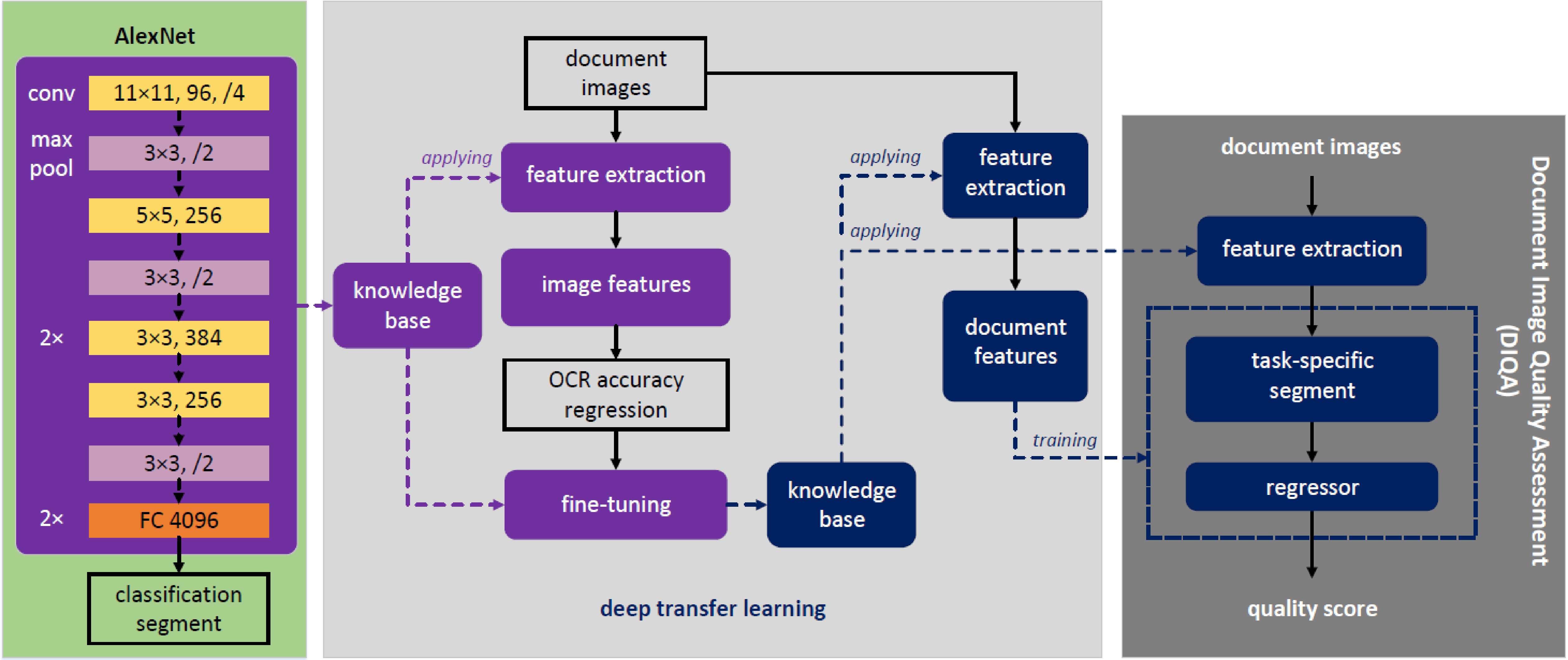
IQA

A Unified Approach to Image Quality Assessment

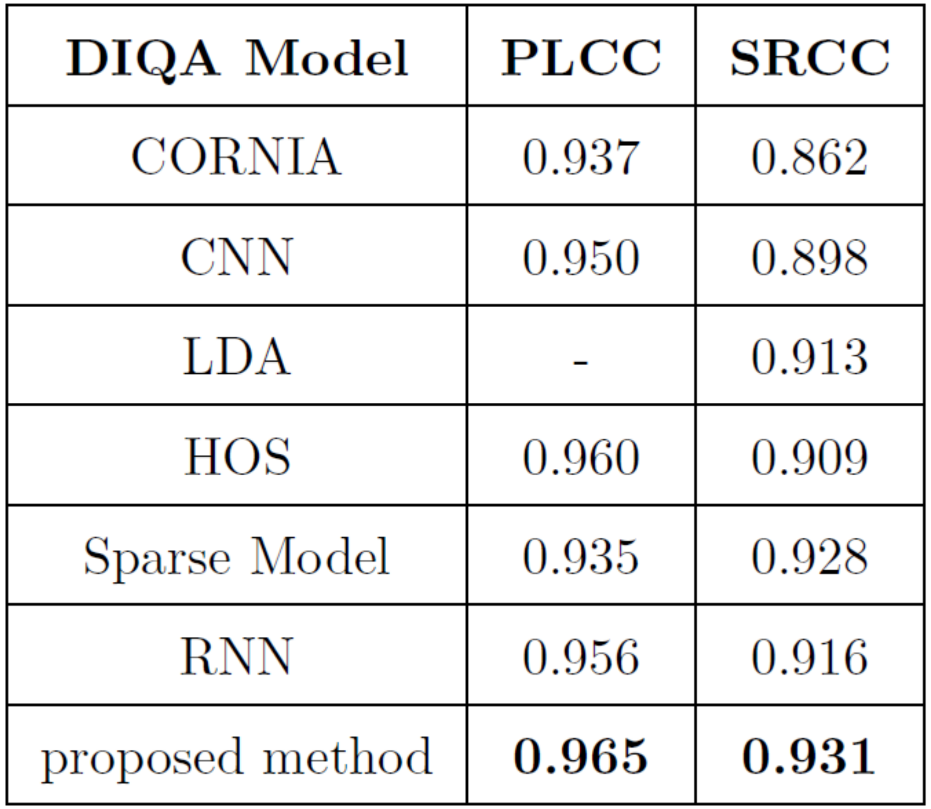


- A unified model to process natural and document images simultaneously
- Content-aware such that different types of quality information is provided according to different types of input images

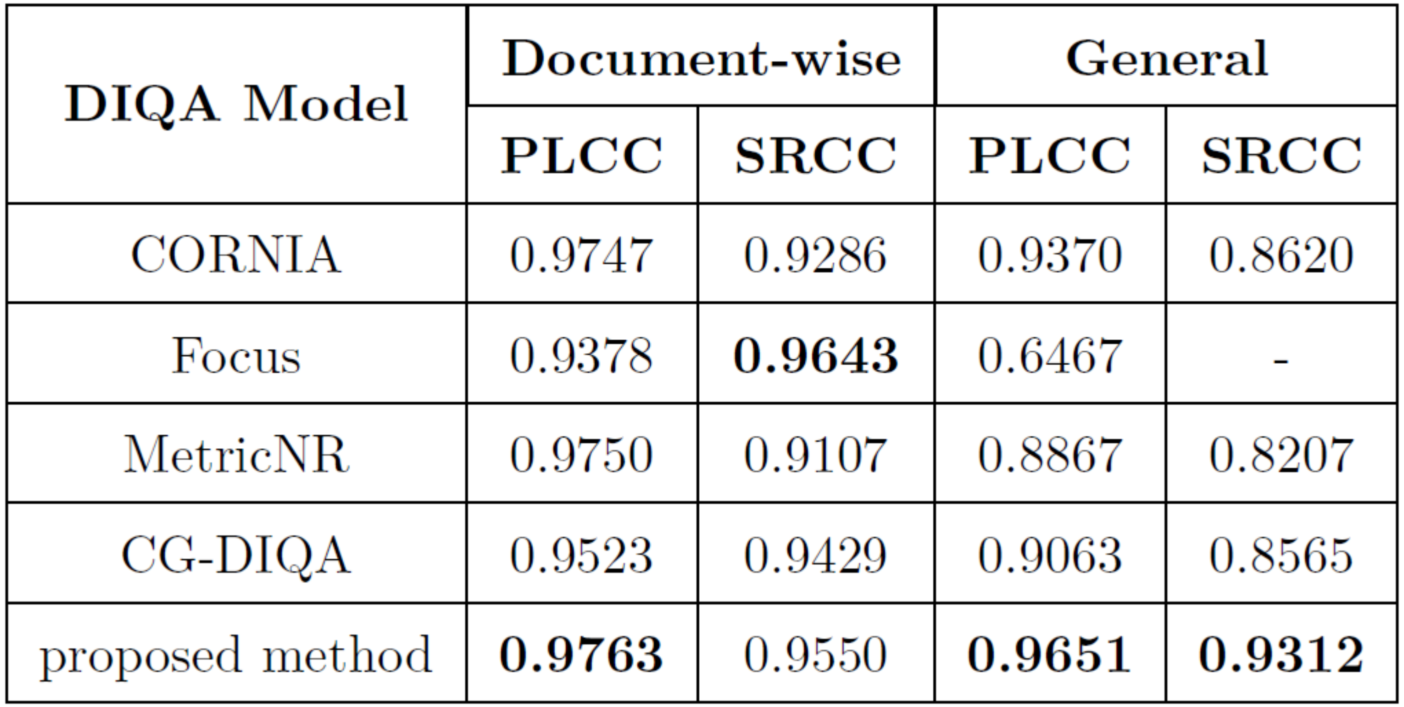
Document Image Quality Assessment based on Transfer Learning



Document Image Quality Assessment based on Transfer Learning



DIQA Model	PLCC	SRCC
CORNIA	0.937	0.862
CNN	0.950	0.898
LDA	-	0.913
HOS	0.960	0.909
Sparse Model	0.935	0.928
RNN	0.956	0.916
proposed method	**0.965**	**0.931**

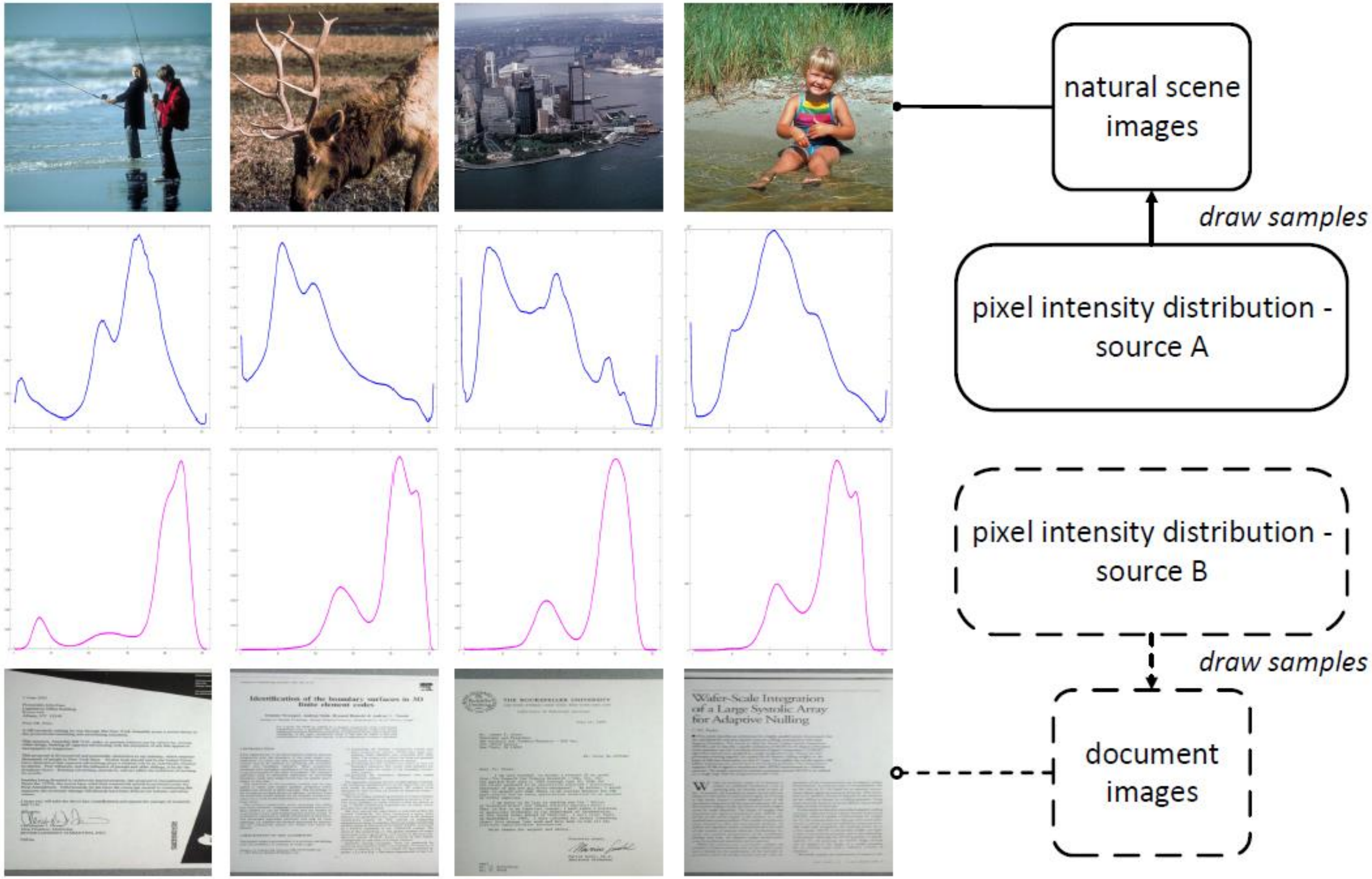


DIQA Model	Document-wise		General	
	PLCC	SRCC	PLCC	SRCC
CORNIA	0.9747	0.9286	0.9370	0.8620
Focus	0.9378	**0.9643**	0.6467	-
MetricNR	0.9750	0.9107	0.8867	0.8207
CG-DIQA	0.9523	0.9429	0.9063	0.8565
proposed method	**0.9763**	0.9550	**0.9651**	**0.9312**

➢ ***Cross-Domain Homogeneity between Natural and Document Images***

The knowledge learned on natural image processing can be effectively exploited for the OCR accuracy prediction of document images.

Unified Image Quality Assessment



natural scene images

draw samples

pixel intensity distribution - source A

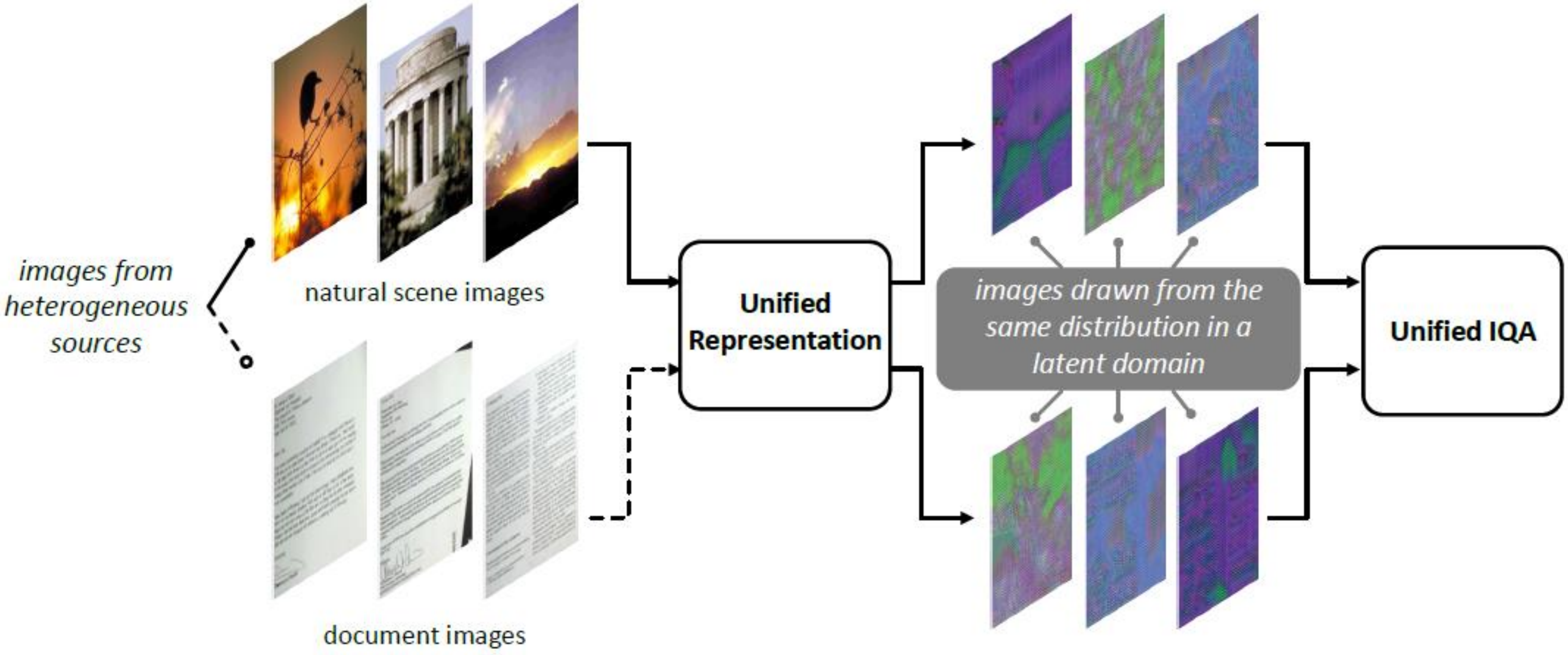
pixel intensity distribution - source B

draw samples

document images

- ***Cross-Domain Homogeneity between Natural and Document Images***
- Possible to process natural and document images simultaneously within one quality assessment model
- Balanced performance on these two types of images can be obtained with the UIQA model
- The process of learning a common representation is mixed with that of regressing the common representation towards respective quality scores – difficult to investigate and develop

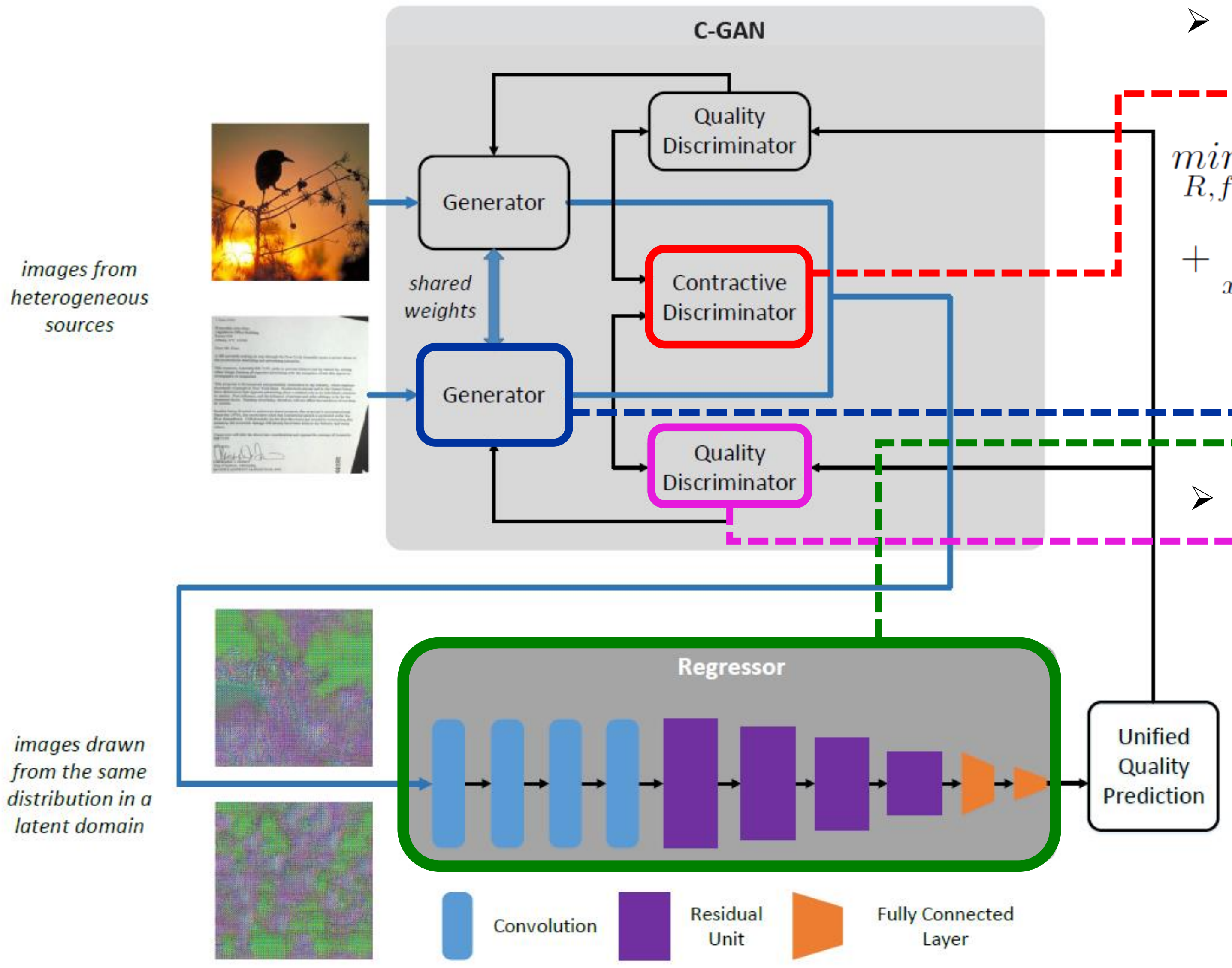
Unified Image Quality Assessment based on Contractive GAN


images from heterogeneous sources
natural scene images
document images
Unified Representation
images drawn from the same distribution in a latent domain
Unified IQA

- ***Cross-Domain Homogeneity between Natural and Document Images***

- Learning a common representation (i.e. a generalization) of natural and document images in a latent domain
- The process of generalization is separated from that of regression
- The quality assessor operates as if it is processing a single type of images

Unified Image Quality Assessment based on Contractive GAN


C-GAN
Quality Discriminator
Generator
Contractive Discriminator
shared weights
Generator
Quality Discriminator
images from heterogeneous sources
images drawn from the same distribution in a latent domain
Regressor
Unified Quality Prediction
Convolution
Residual Unit
Fully Connected Layer

- ***Main Objective:***

$$\min_{R,f}\ \max_{D}\Big\{\mathbb{E}_{x\sim p_B}\{\log[\boxed{D}(f(x))]\} + \mathbb{E}_{x\sim p_A}\{\log[1-D(f(x))]\} + \mathbb{E}_{x\sim p_A}\{|R(\boxed{f}(x)) - t_A|\} + \mathbb{E}_{x\sim p_B}\{|\boxed{R}(f(x)) - t_B|\}\Big\}$$

- ***Quality Discriminator:***

$$\max_{D_A}\Big\{\mathbb{E}_{x\sim p_A}\{\log[1-|\boxed{D_A}(f(x))-\sigma(x)|]\}\Big\}+$$

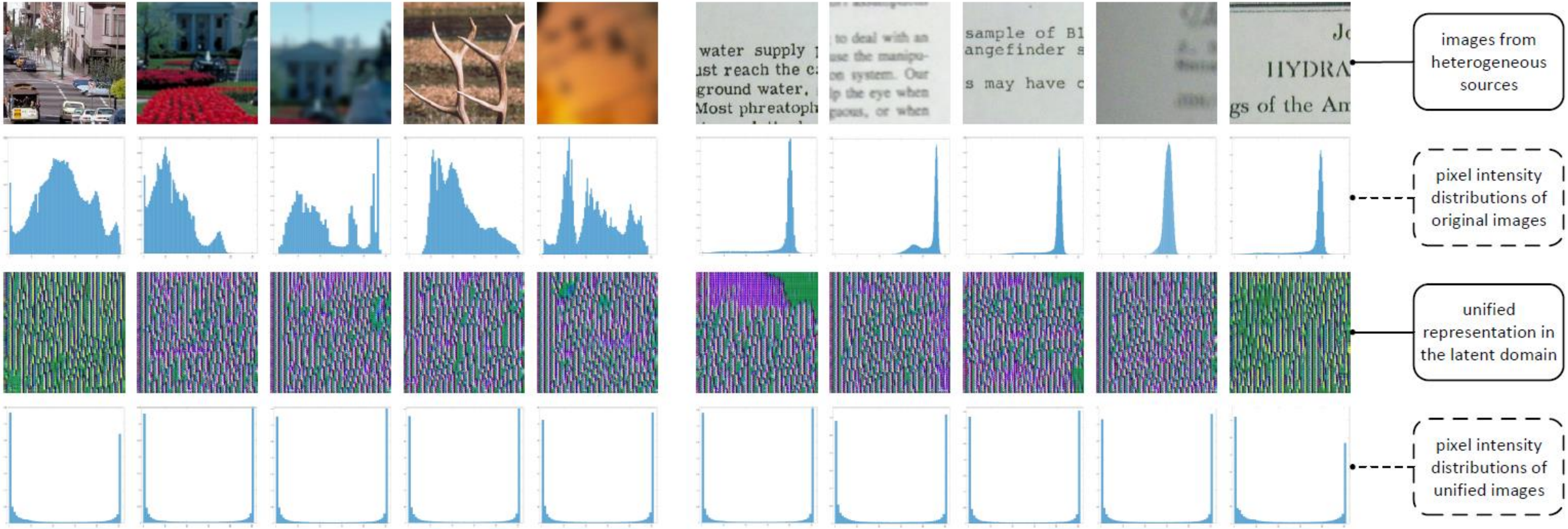
$$\max_{D_B}\Big\{\mathbb{E}_{x\sim p_B}\{\log[1-|D_B(f(x))-\sigma(x)|]\}\Big\}$$

where:

$$\sigma(x) = \begin{cases} 1, & \text{if } |R(x)-t| \le \epsilon \\ 0, & \text{otherwise} \end{cases}$$

Unified Image Quality Assessment based on Contractive GAN

- Qualitative evaluation: visualization of the operation of the C-GAN model



images from heterogeneous sources

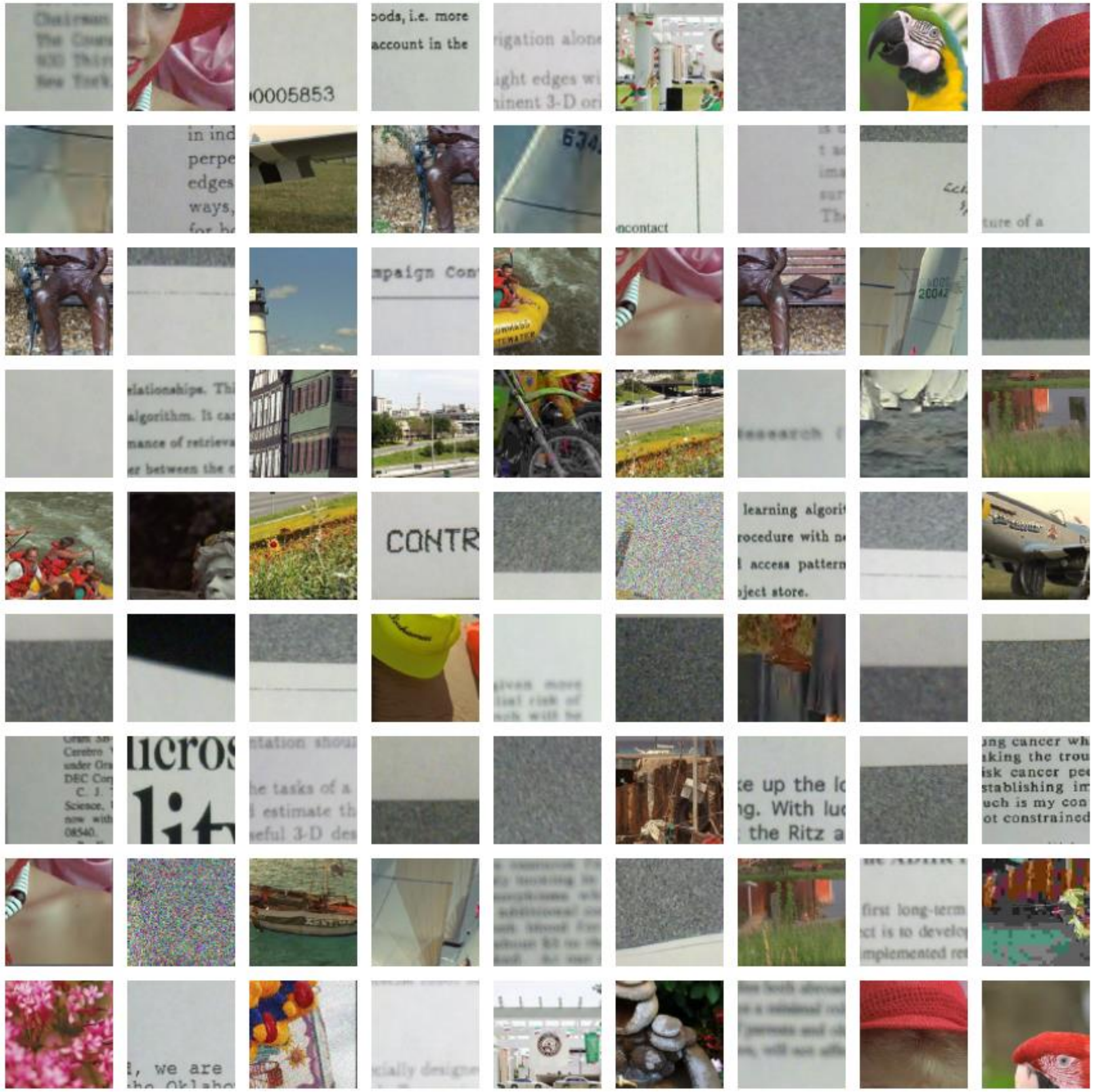
pixel intensity distributions of original images

unified representation in the latent domain

pixel intensity distributions of unified images

Unified Image Quality Assessment based on Transfer Learning

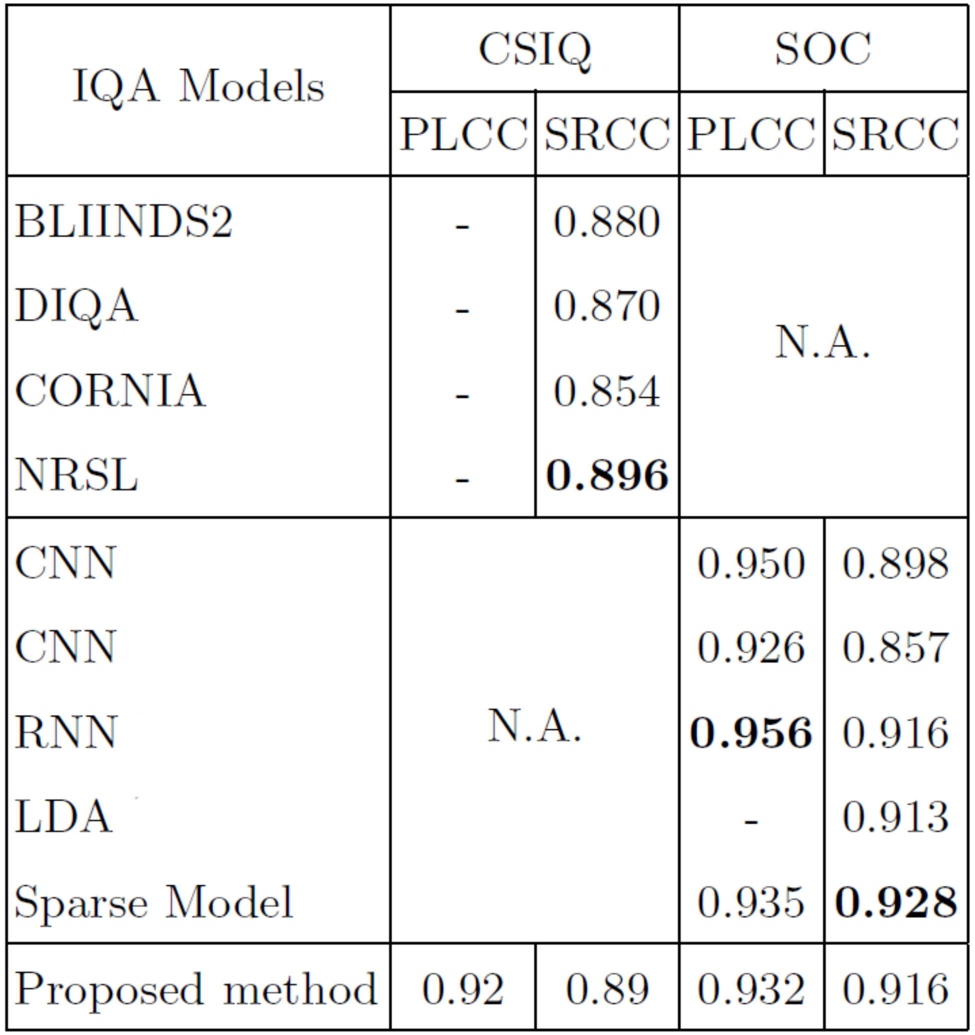
LIVE + SOC



CSIQ + SOC

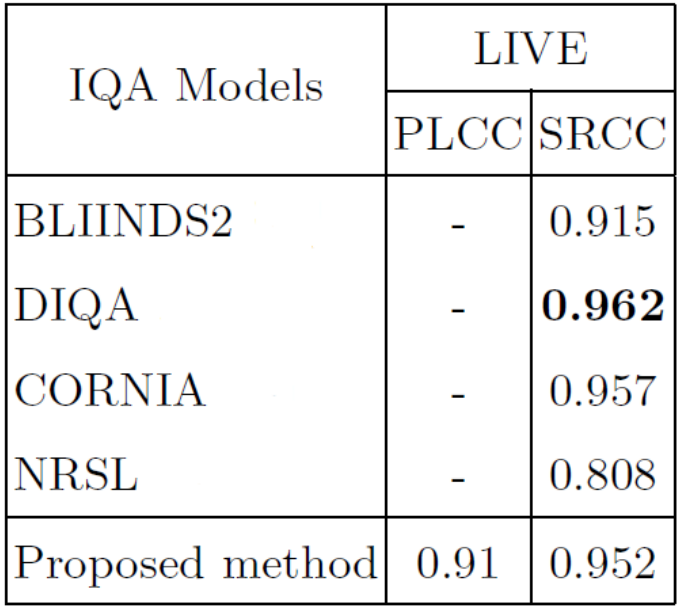
Unified Image Quality Assessment based on Contractive GAN

- Comparing to content-specific IQA and DIQA models



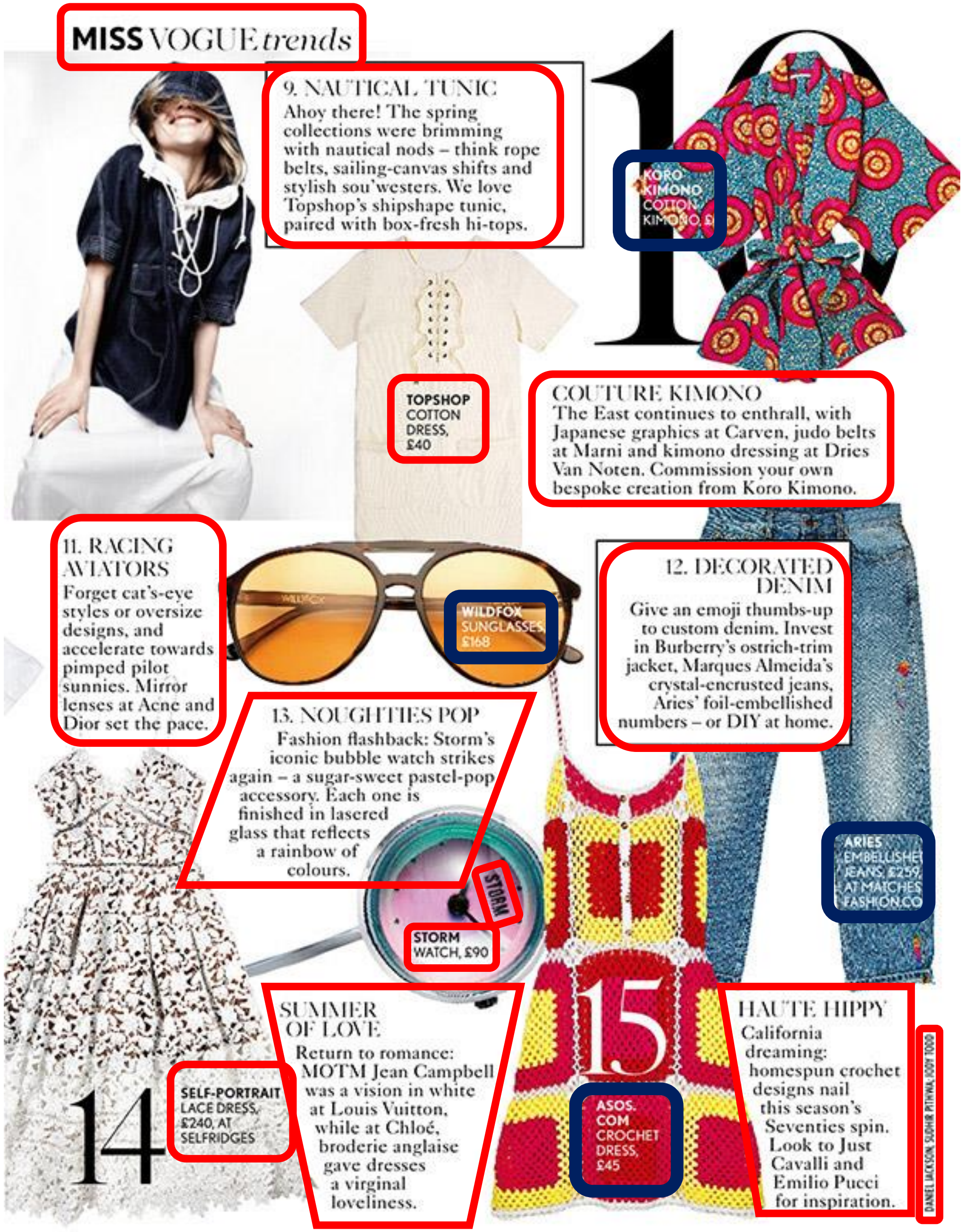
IQA Models	CSIQ		SOC	
	PLCC	SRCC	PLCC	SRCC
BLIINDS2	-	0.880	N.A.	
DIQA	-	0.870		
CORNIA	-	0.854		
NRSL	-	**0.896**		
CNN	N.A.		0.950	0.898
CNN			0.926	0.857
RNN			**0.956**	0.916
LDA			-	0.913
Sparse Model			0.935	**0.928**
Proposed method	0.92	0.89	0.932	0.916

- Cross-dataset evaluation of the proposed UIQA model on the natural scene images



IQA Models	LIVE	
	PLCC	SRCC
BLIINDS2	-	0.915
DIQA	-	**0.962**
CORNIA	-	0.957
NRSL	-	0.808
Proposed method	0.91	0.952

Gestalt Principles and Text Homogeneity



MISS VOGUE *trends*

9. NAUTICAL TUNIC
Ahoy there! The spring collections were brimming with nautical nods – think rope belts, sailing-canvas shifts and stylish sou'westers. We love Topshop's shipshape tunic, paired with box-fresh hi-tops.

KORO KIMONO, COTTON KIMONO, £

TOPSHOP COTTON DRESS, £40

10

COUTURE KIMONO
The East continues to enthrall, with Japanese graphics at Carven, judo belts at Marni and kimono dressing at Dries Van Noten. Commission your own bespoke creation from Koro Kimono.

11. RACING AVIATORS
Forget cat's-eye styles or oversize designs, and accelerate towards pimped pilot sunnies. Mirror lenses at Acne and Dior set the pace.

WILDFOX SUNGLASSES £168

12. DECORATED DENIM
Give an emoji thumbs-up to custom denim. Invest in Burberry's ostrich-trim jacket, Marques Almeida's crystal-encrusted jeans, Aries' foil-embellished numbers – or DIY at home.

13. NOUGHTIES POP
Fashion flashback: Storm's iconic bubble watch strikes again – a sugar-sweet pastel-pop accessory. Each one is finished in lasered glass that reflects a rainbow of colours.

STORM

STORM WATCH, £90

ARIES EMBELLISHED JEANS, £259, AT MATCHESFASHION.COM

14

SUMMER OF LOVE
Return to romance: MOTM Jean Campbell was a vision in white at Louis Vuitton, while at Chloé, broderie anglaise gave dresses a virginal loveliness.

SELF-PORTRAIT LACE DRESS, £240, AT SELFRIDGES

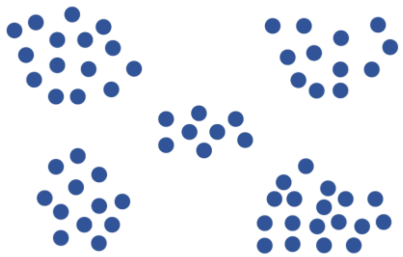
15

ASOS.COM CROCHET DRESS, £45

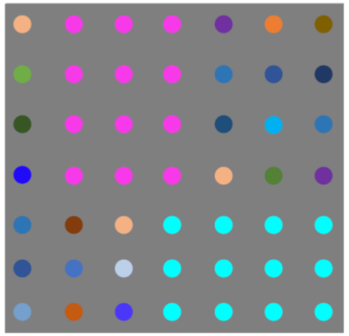
HAUTE HIPPY
California dreaming: homespun crochet designs nail this season's Seventies spin. Look to Just Cavalli and Emilio Pucci for inspiration.

DANIEL JACKSON; SUDHIR PITHWA; JODY TODD

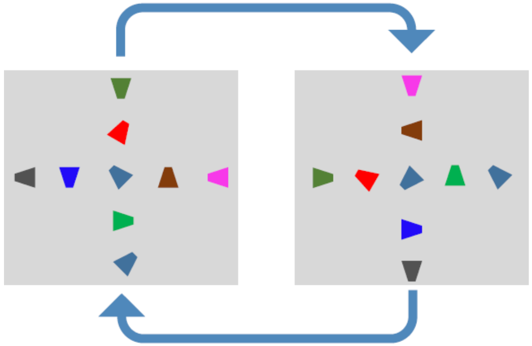
- ***Proximity***



- ***Similarity***

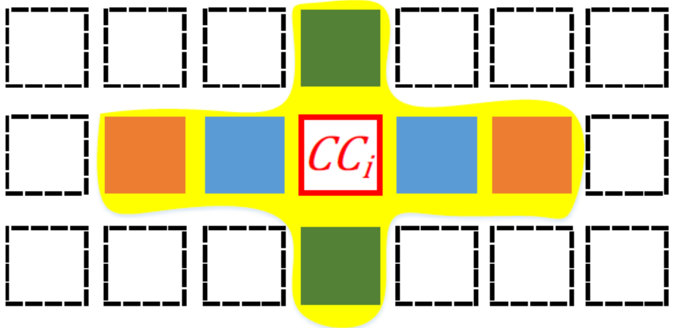


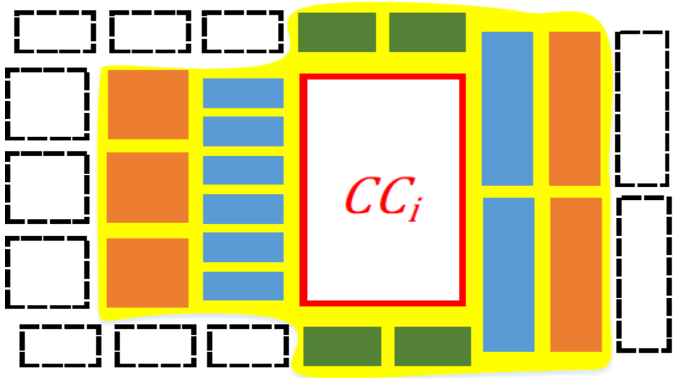
- ***Symmetry***



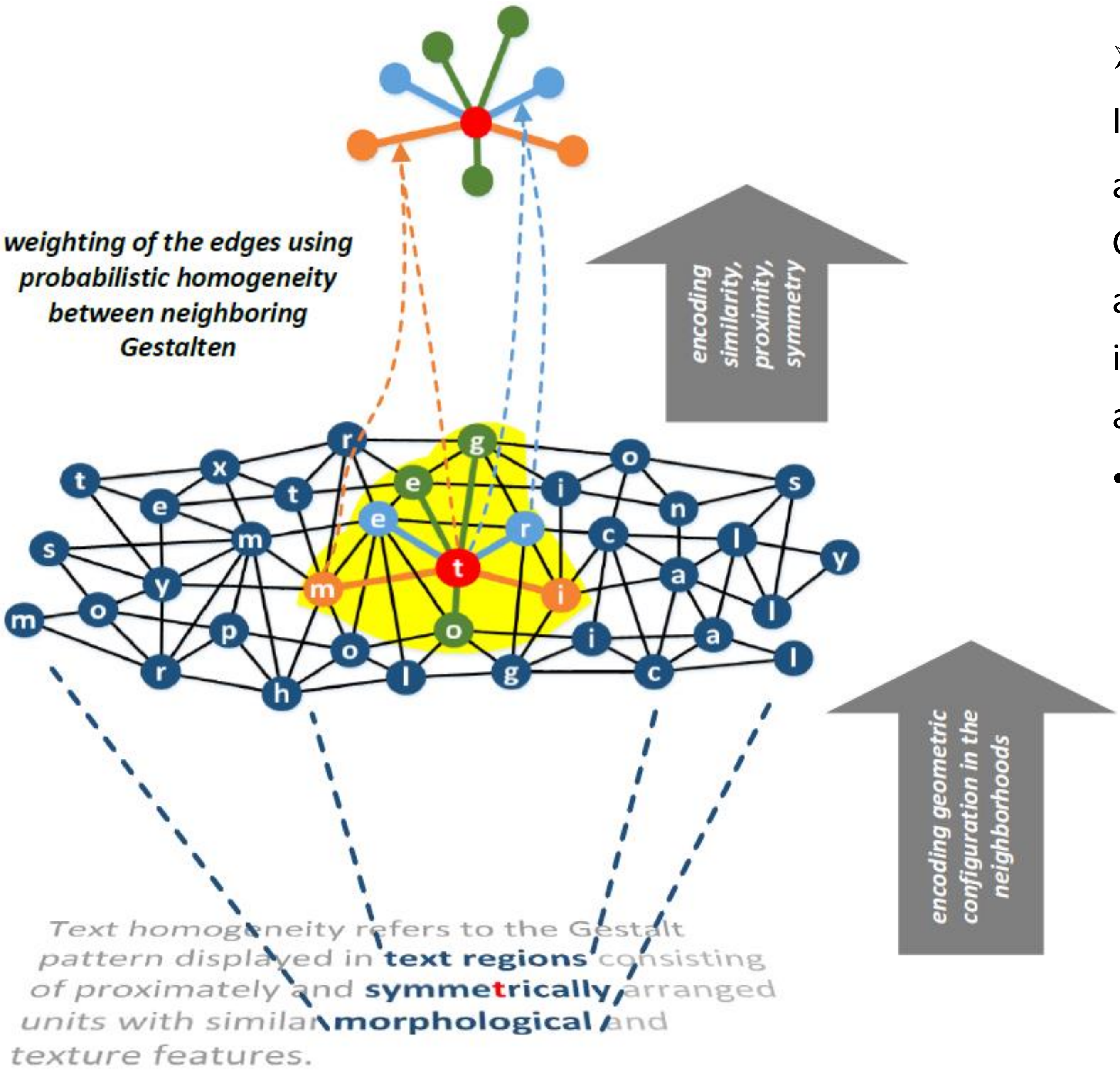
- ***Conceptualization***

Text homogeneity is the homogeneous pattern displayed in text regions, which consists of **proximately** and **symmetrically** arranged units with **similar** ***morphological*** and ***texture*** features.





Probabilistic Local Text Homogeneity – A Neighborhood Graph



weighting of the edges using probabilistic homogeneity between neighboring Gestalten

encoding similarity, proximity, symmetry

encoding geometric configuration in the neighborhoods

Text homogeneity refers to the Gestalt pattern displayed in **text regions** consisting of proximately and **symmetrically** arranged units with similar **morphological** and texture features.

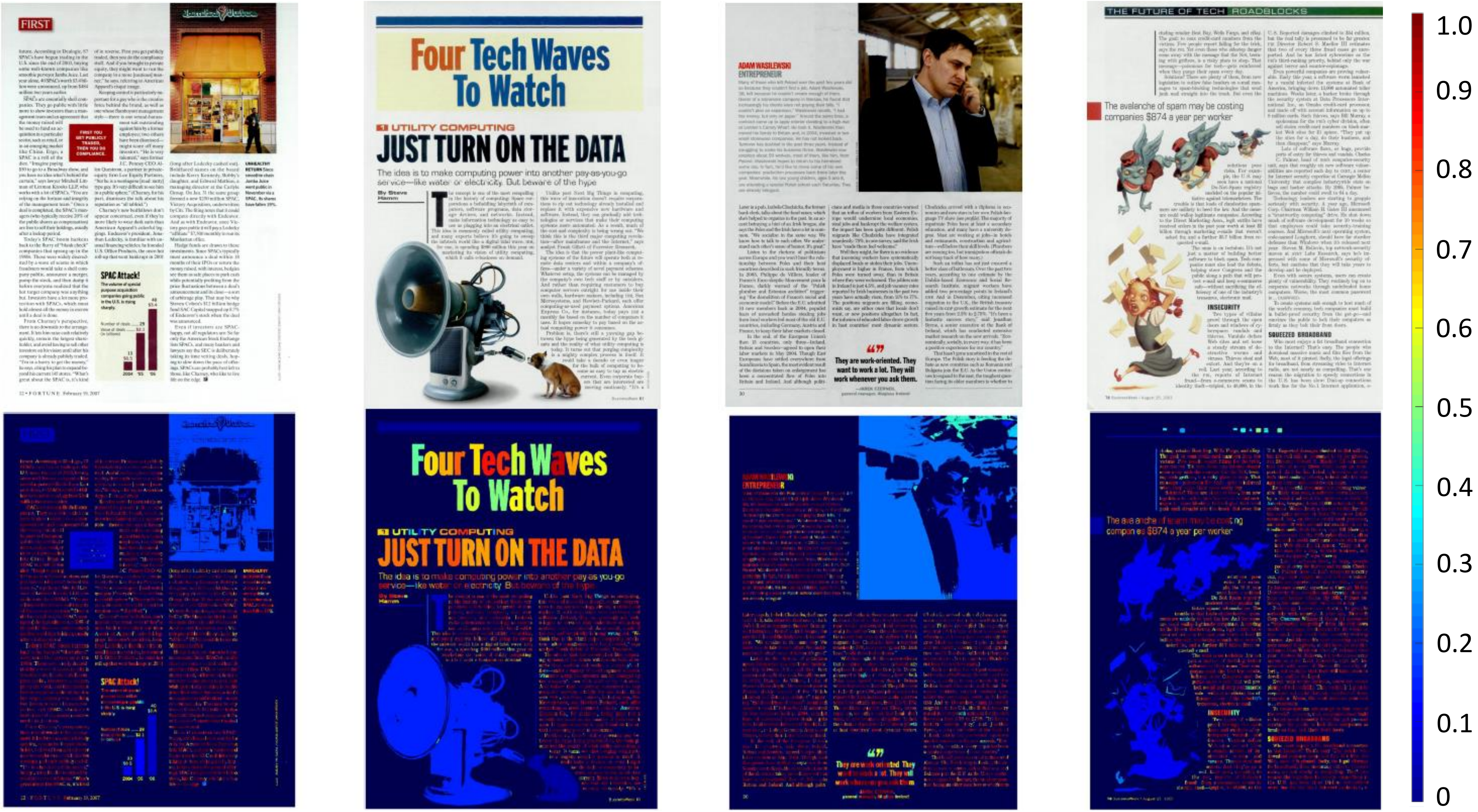
- ***Description of local text homogeneity on G(V,E)***

If we take a one-step walk from a Gestalt CC_i by following an arbitrary (**symmetry**) direction, and arrives at another Gestalt, say CC_j, the probability that CC_j is located within a short (**proximity**) distance and resembles (**similarity**) CC_i is higher when CC_j is a text component (e.g. a letter from a paragraph).

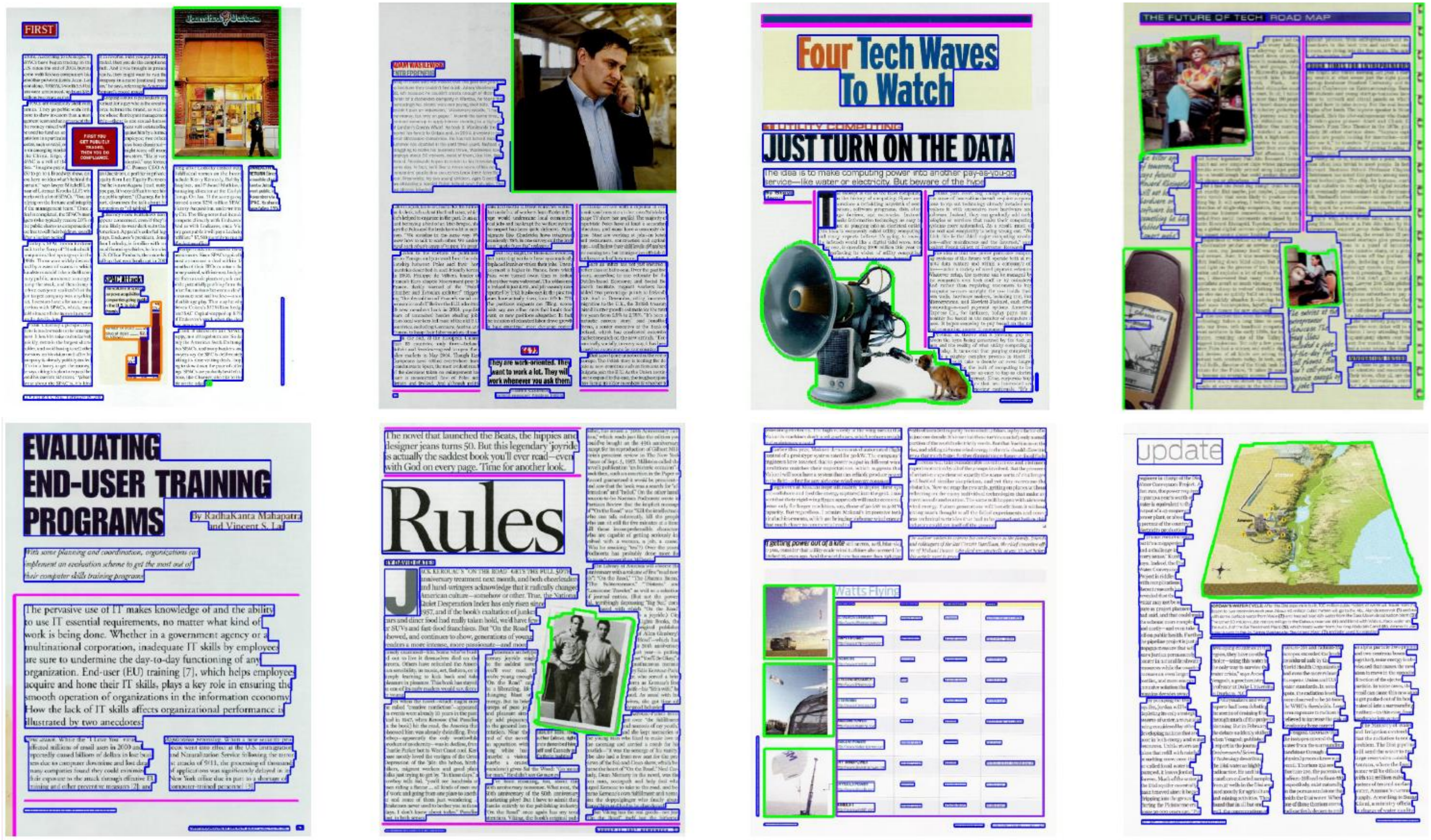
- probabilistic weighting $w_{ij} = P(S_{ij} = s_{ij}^+)$

$$S_{ij} = \begin{cases} s_{ij}^+, & \text{if } CC_i \text{ and } CC_j \text{ are homogeneous,} \\ s_{ij}^-, & \text{if } CC_i \text{ and } CC_j \text{ are heterogeneous;} \end{cases}$$

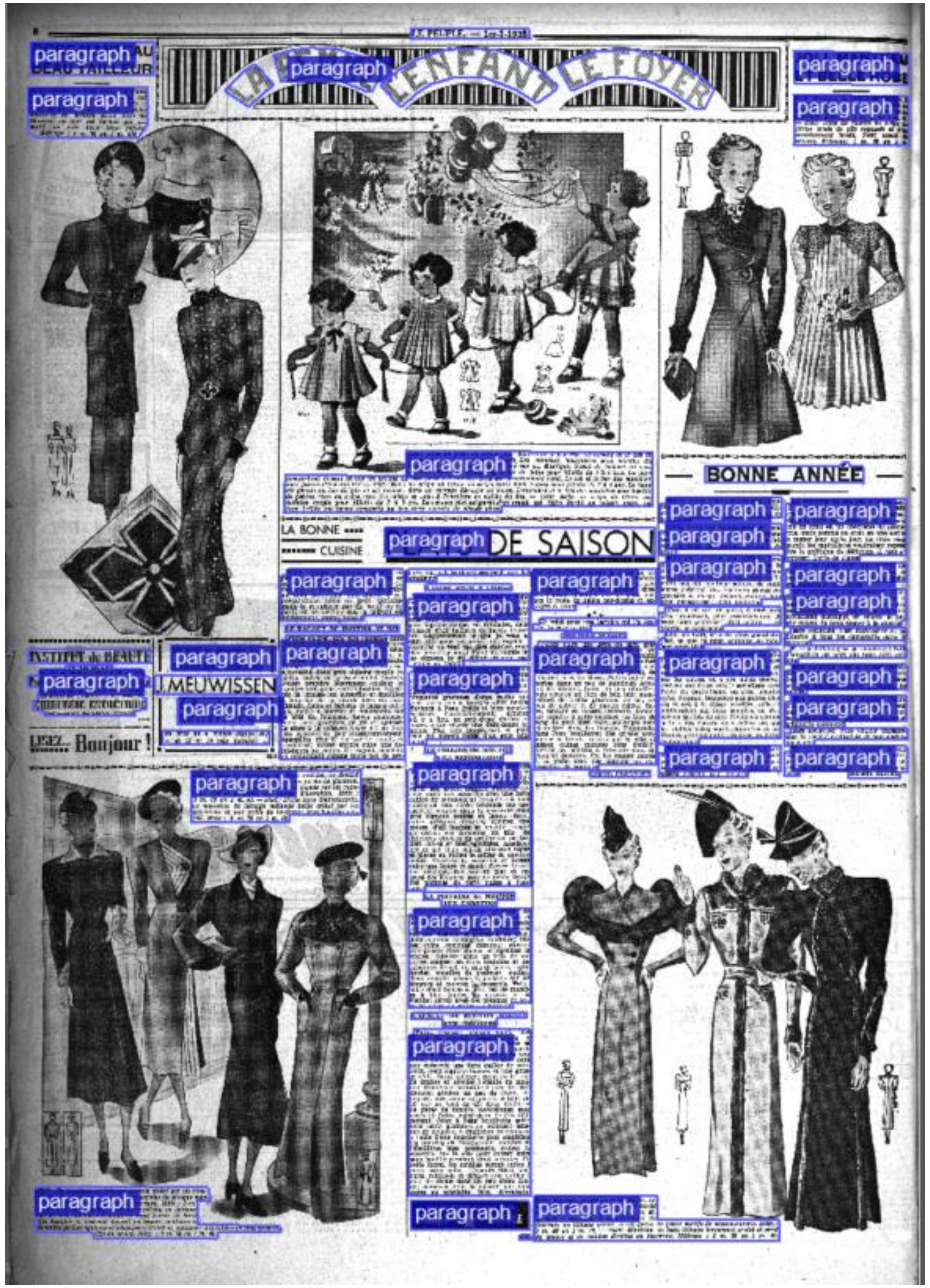
Probabilistic Local Text Homogeneity – Resultant Probability Map


FIRST
SPAC Attack!
Four Tech Waves To Watch
UTILITY COMPUTING
JUST TURN ON THE DATA
The idea is to make computing power into another pay-as-you-go service—like water or electricity. But beware of the hype
ADAM WASILEWSKI
ENTREPRENEUR
They are work-oriented. They want to work a lot. They will work whenever you ask them.
THE FUTURE OF TECH ROADBLOCKS
The avalanche of spam may be costing companies $874 a year per worker
INSECURITY
SQUEEZED BROADBAND
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

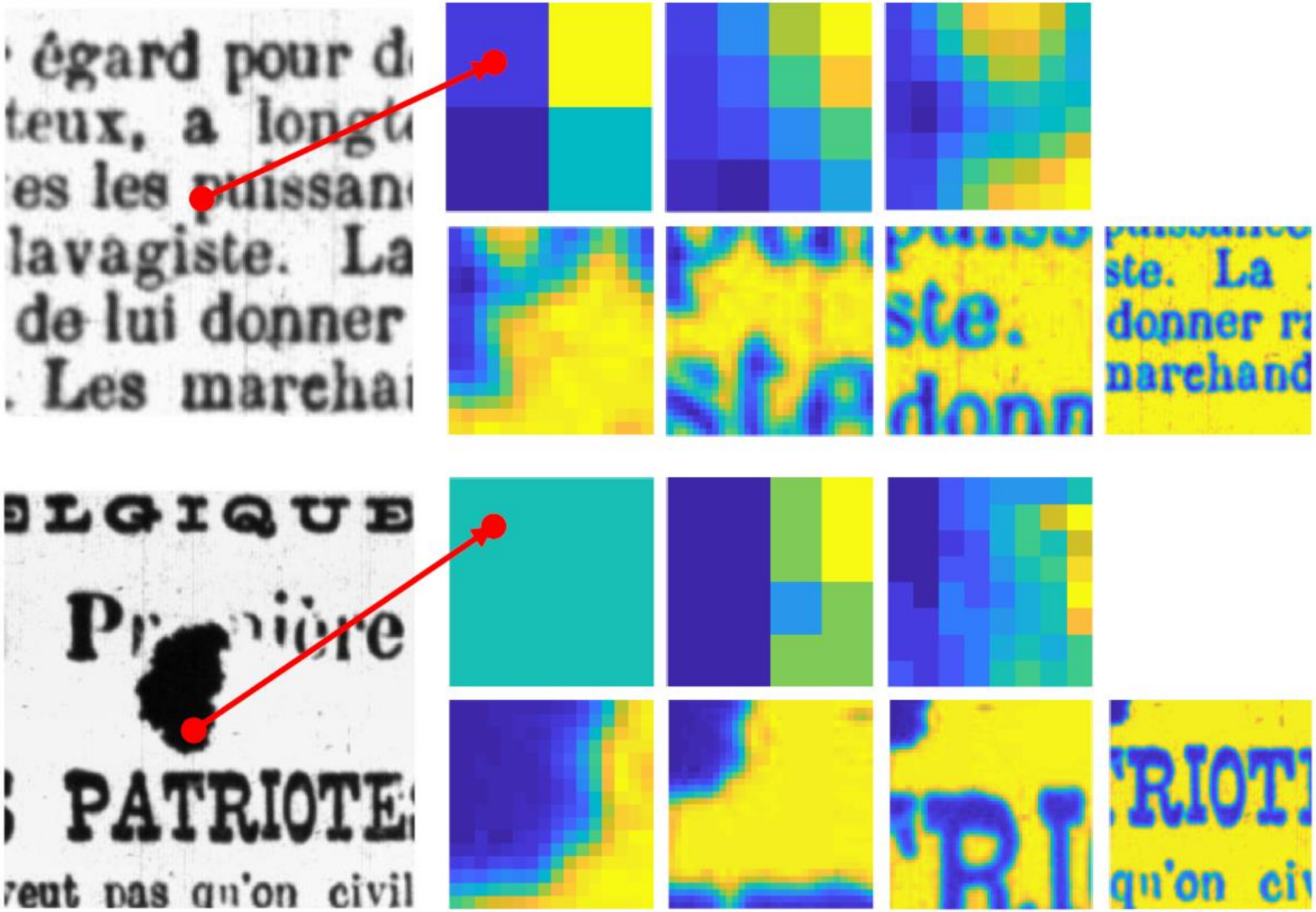
Document Segmentation with Probabilistic Homogeneity - Performances


FIRST
ADAM WASILEWSKI
ENTREPRENEUR
They are work-oriented. They want to work a lot. They will work whenever you ask them
Four Tech Waves To Watch
UTILITY COMPUTING
JUST TURN ON THE DATA
The idea is to make computing power into another pay-as-you-go service—like water or electricity. But beware of the hype
THE FUTURE OF TECH ROAD MAP
EVALUATING END-USER TRAINING PROGRAMS
By RadhaKanta Mahapatra and Vincent S. Lai
With some planning and coordination, organizations can implement an evaluation scheme to get the most out of their computer skills training programs
The pervasive use of IT makes knowledge of and the ability to use IT essential requirements, no matter what kind of work is being done. Whether in a government agency or a multinational corporation, inadequate IT skills by employees are sure to undermine the day-to-day functioning of any organization. End-user (EU) training [7], which helps employees acquire and hone their IT skills, plays a key role in ensuring the smooth operation of organizations in the information economy. How the lack of IT skills affects organizational performance is illustrated by two anecdotes:
The novel that launched the Beats, the hippies and designer jeans turns 50. But this legendary 'joyride' is actually the saddest book you'll ever read—even with God on every page. Time for another look.
Rules
BY DAVID GATES
Watts Flying
Update

Document Segmentation with Probabilistic Homogeneity - Performances

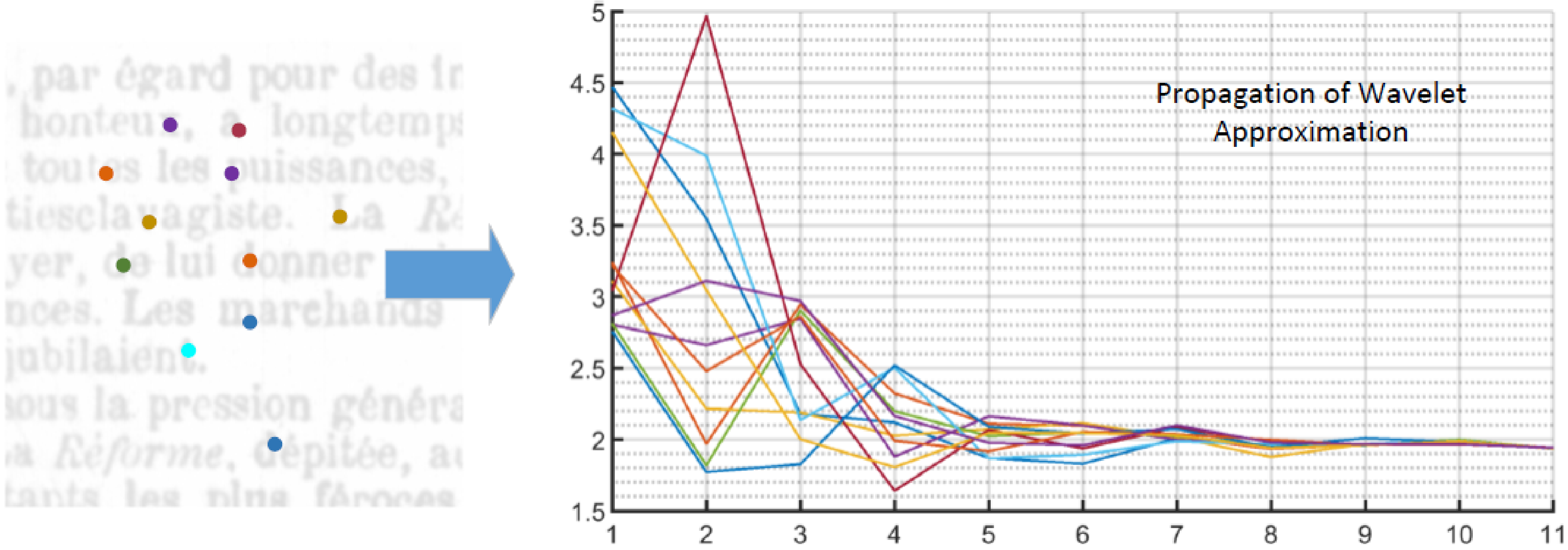


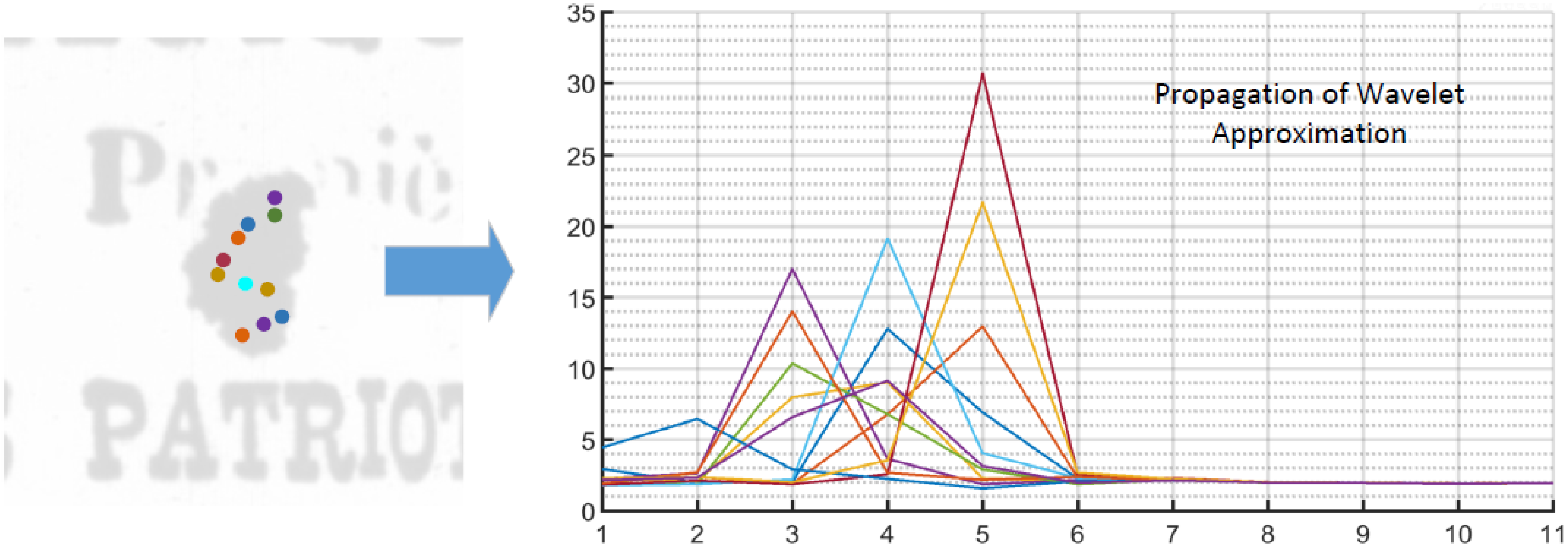
Propagation of Wavelet Approximation



- ***Text Homogeneity Revisit***
- Text Homogeneity pattern
- Neighborhood transition

Propagation of Wavelet Approximation





➢ ***Wavelet Propagation***

propagation of wavelet approximation (PWA) and propagation of cone-of-influence wavelet approximation (PCWA).

- PWA

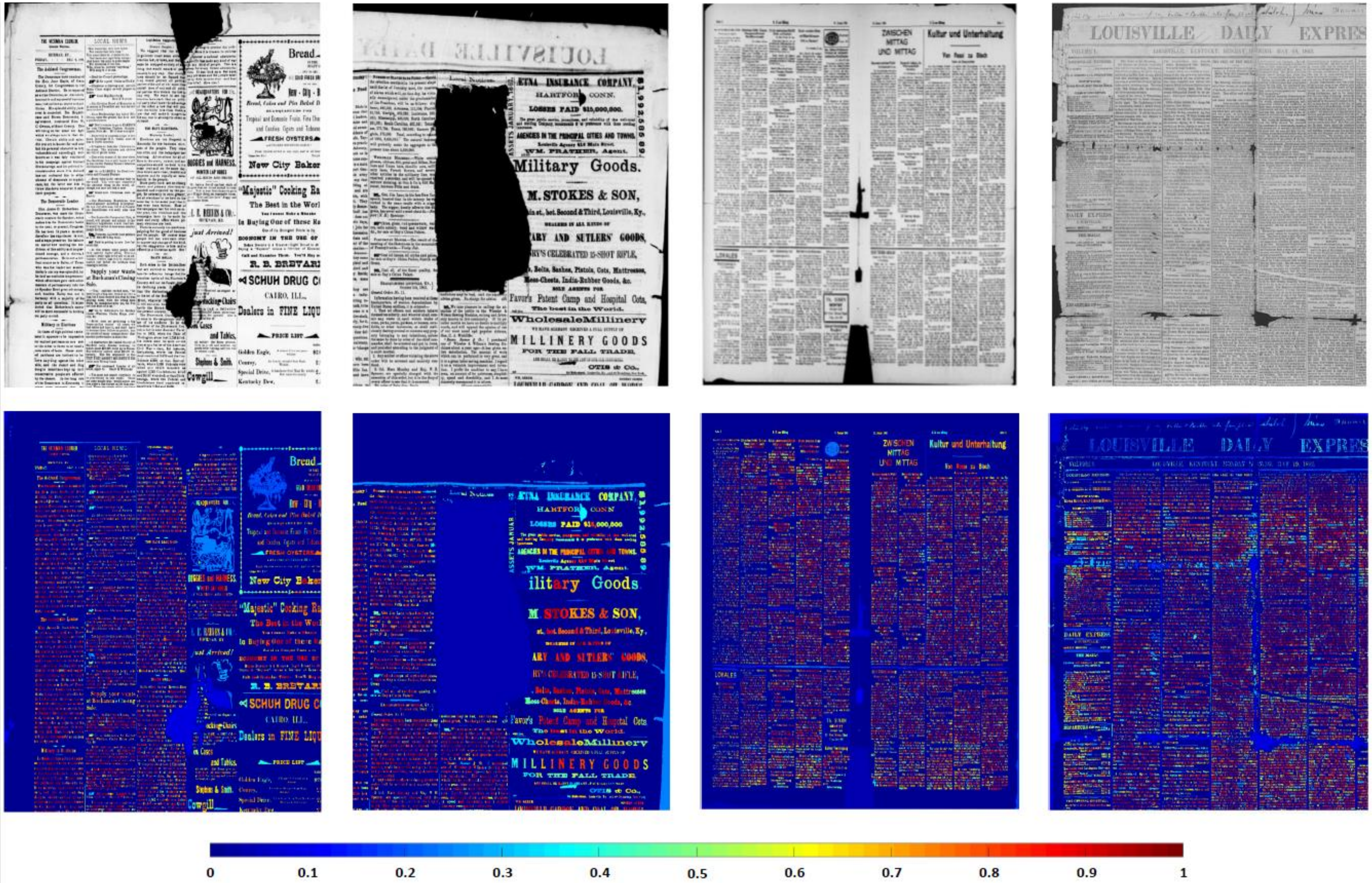
$$\alpha_{n\to k,l} \triangleq \log_2\left(\frac{1}{k-n}\sum_{j=n}^{k-1}\frac{|w_\phi^{j+1,l}|}{|w_\phi^{j,l}|}\right)$$

- PCWA

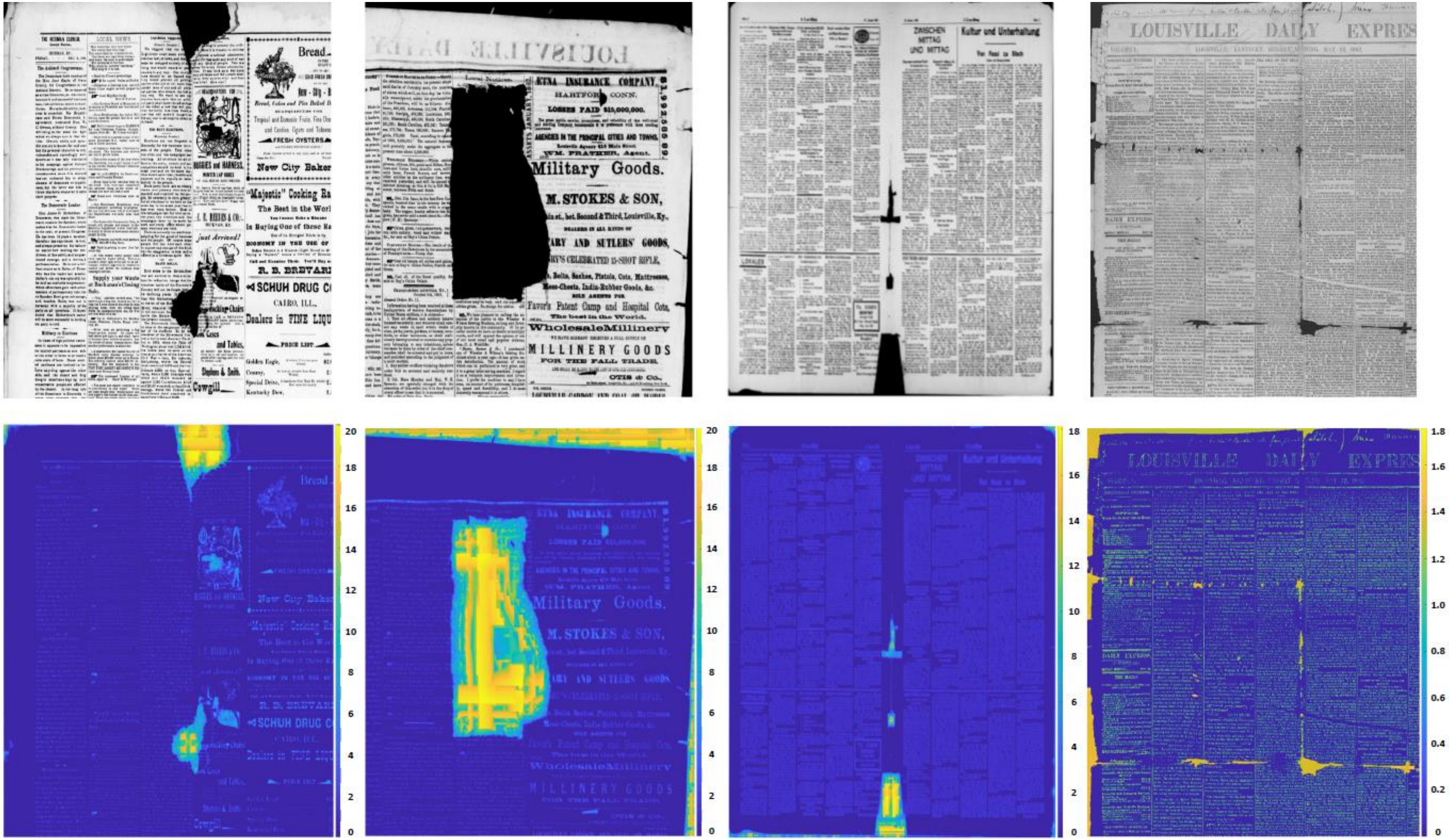
$$\beta_{n\to k,l} \triangleq \log_2\left(\frac{1}{k-n}\sum_{j=n}^{k-1}\frac{|I_{j+1,l}|}{|I_{j,l}|}\right),$$

$$I_{j,l} \triangleq \sum_{m\in C(j,l)} |w_\phi^{j,m}|$$

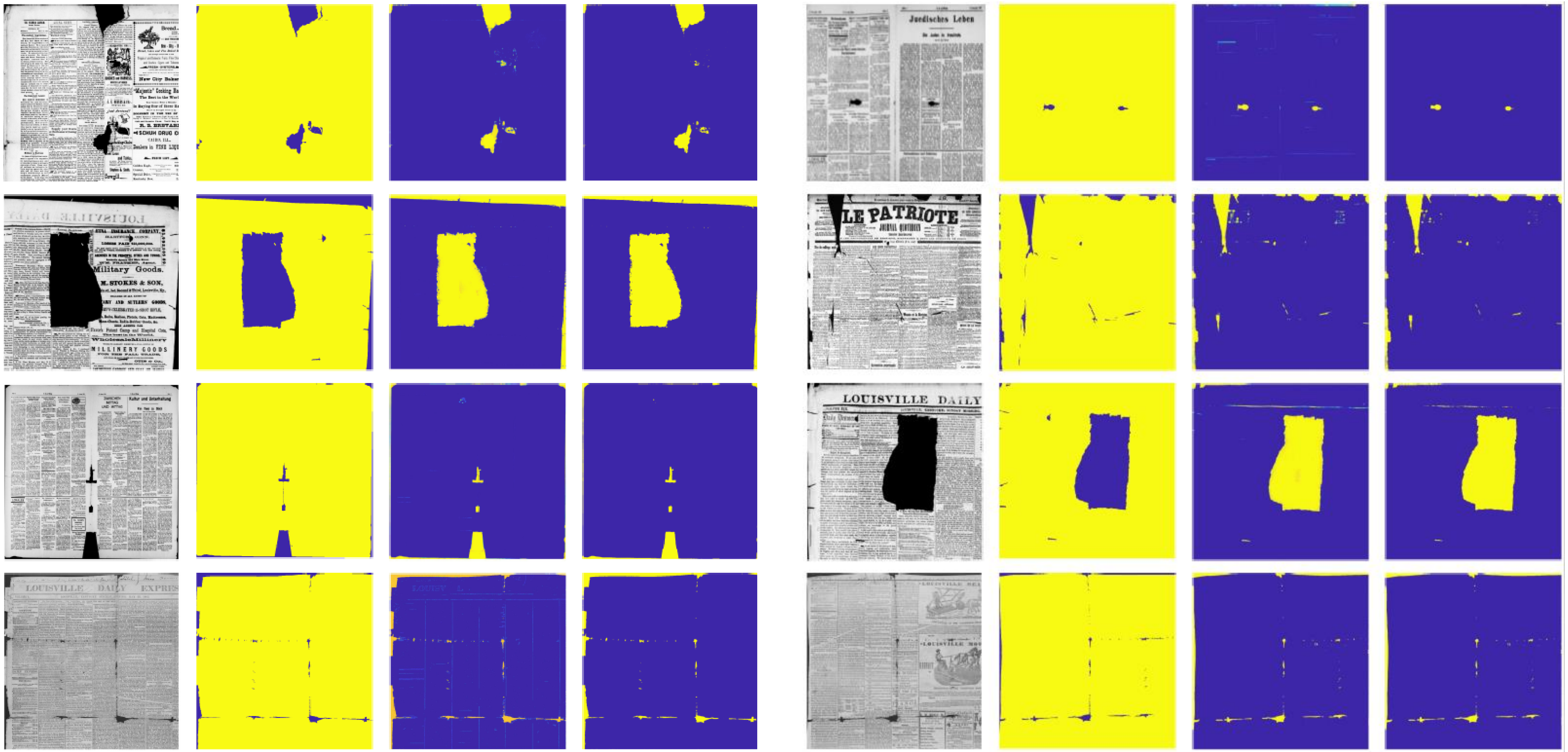
Bayesian Distortion Recognition



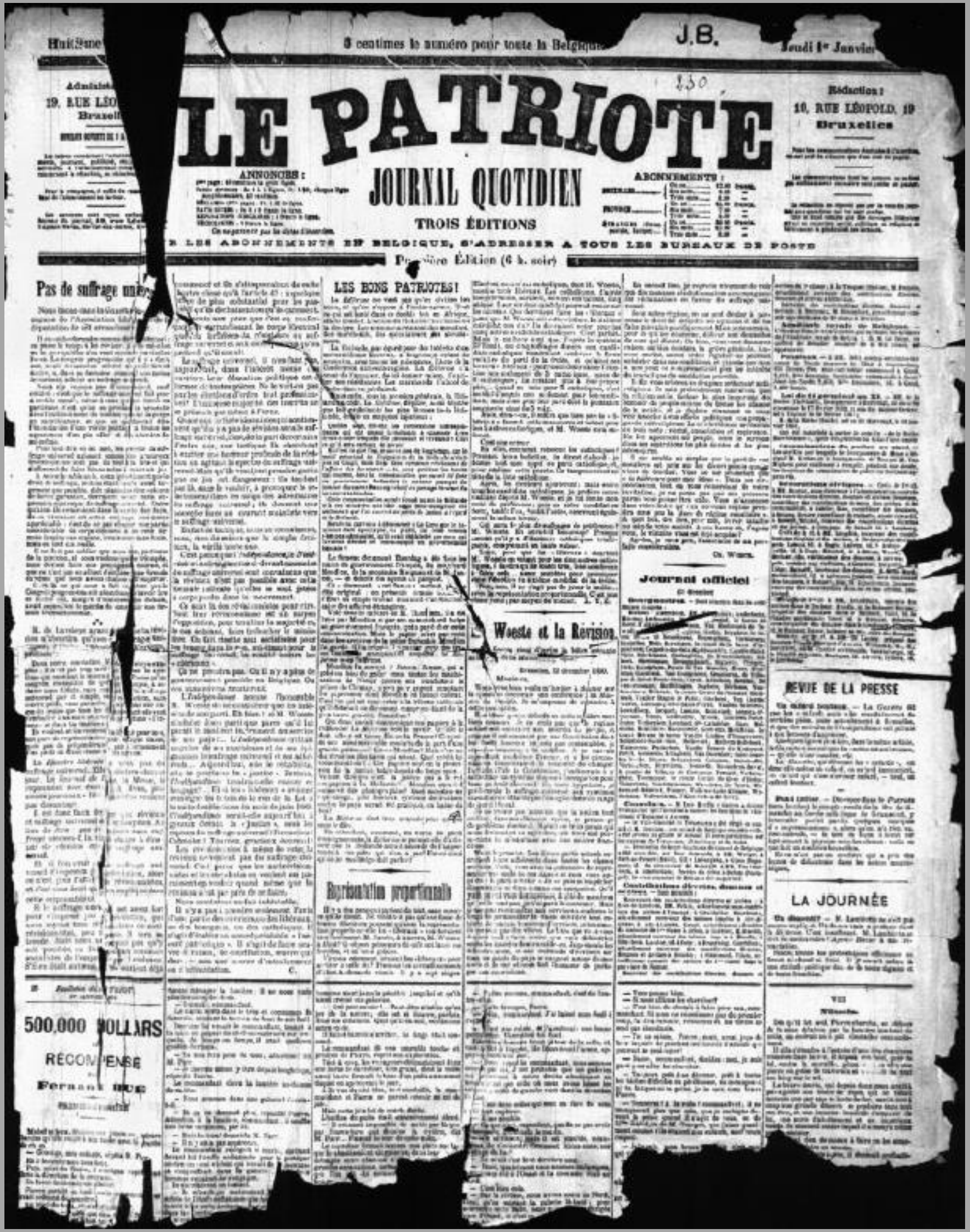
Bayesian Distortion Recognition



Bayesian Distortion Recognition



Conclusion



Huitième … 5 centimes le numéro pour toute la Belgique … J.B. … Jeudi 1er Janvier

Administ… 19, RUE LÉO… Bruxell…

LE PATRIOTE

JOURNAL QUOTIDIEN

TROIS ÉDITIONS

ANNONCES :

ABONNEMENTS :

Rédaction : 19, RUE LÉOPOLD, 19 Bruxelles

…R LES ABONNEMENTS EN BELGIQUE, S'ADRESSER A TOUS LES BUREAUX DE POSTE

Pr…ière Édition (6 h. soir)

Pas de suffrage unive…

LES BONS PATRIOTES!

Woeste et la Révision

Représentation proportionnelle

Journal officiel

REVUE DE LA PRESSE

LA JOURNÉE

500,000 DOLLARS

RÉCOMPENSE

Q & A

Thank you for your attention!

Q & A