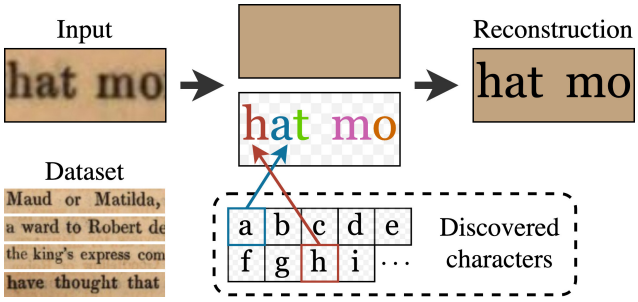
The Learnable Typewriter: A Generative Approach to Text Analysis

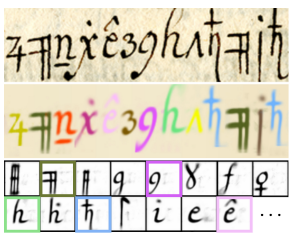
Ioannis Siglidis Nicolas Gonthier Julien Gaubil Tom Monnier Mathieu Aubry
LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS, Marne-la-Vallée, France
https://imagine.enpc.fr/~siglidii/learnable-typewriter

Abstract

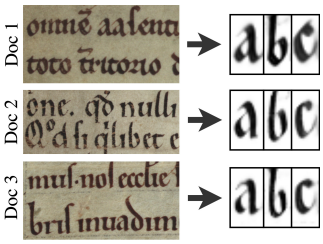
We present a generative document-specific approach to character analysis and recognition in text lines. Our main idea is to build on unsupervised multi-object segmentation methods and in particular those that reconstruct images based on a limited amount of visual elements, called sprites. Taking as input a set of text lines with similar font or handwriting, our approach can learn a large number of different characters and leverage line-level annotations when available. Our contribution is twofold. First, we provide the first adaptation and evaluation of a deep unsupervised multi-object segmentation approach for text line analysis. Since these methods have mainly been evaluated on synthetic data in a completely unsupervised setting, demonstrating that they can be adapted and quantitatively evaluated on real images of text and that they can be trained using weak supervision are significant progresses. Second, we show the potential of our method for new applications, more specifically in the field of paleography, which studies the history and variations of handwriting, and for cipher analysis. We demonstrate our approach on three very different datasets: a printed volume of the Google1000 dataset [45, 19], the Copiale cipher [2, 27] and historical handwritten charters from the 12th and early 13th century [6, 43].



(a) The Learnable Typewriter idea



(b) Cipher analysis [27]



(c) Paleographic analysis [43]

Figure 1: **The Learnable Typewriter. (a)** Given a text line dataset, we learn to reconstruct images to discover the underlying characters. Such a generative approach can be used to analyze complex ciphers **(b)** and can be used as an automatic tool to help the study of handwriting variations in historical documents **(c)**.

1. Introduction

A popular approach to document analysis in the 1990s was to learn document-specific character prototypes, which enabled Optical Character Recognition (OCR) [28, 29, 46, 1] but also other applications, such as font classification [21] or document image compression and rendering [38]. This idea culminated in 2013, with the Ocular system [3] which proposed a generative model for printed text lines inspired by the printing process and held the promise of achieving a complete explanation of their appearance. These document-specific generative approaches were however overshadowed by discriminative approaches, whose sole purpose is to perform predictions, and which lead to higher performance at the cost of interpretability, e.g. [16, 33]. In this paper, we explore how modern deep approaches enable revisiting and extending model-based approaches to text line analysis. In particular, we demonstrate an approach that can deal with challenging examples of handwritten documents, opening a new perspective for the study of historical handwriting, paleography.

While discriminative approaches are largely dominant in today's deep learning-based computer vision, a recent set of works revisited generative approaches for unsupervised multi-object object segmentation [5, 10, 18, 17, 47, 7, 9, 11, 23, 41, 37]. Most of them provide results on synthetic data or simple real images [37], and sometimes show qualitative results on simple printed text images [41, 40].

Surprisingly, images of handwritten characters, which were notoriously used in the development of convolutional neural networks [31, 32] and generative adversarial networks [14] were largely overlooked by these approaches.

We build on recent sprite-based unsupervised image decomposition approaches [41, 37] that provide an interpretable decomposition of images into a dictionary of visual elements, referred to as sprites. These methods jointly optimize both the sprites and the neural networks that predict their position and color. Intuitively, we would like to adapt these methods so that, from text lines that are extracted from any given document, they could learn sprites that correspond to each character. By adapting MarioNette [41] to perform text line analysis, we provide quantitative evaluation on real data and an analysis of the limitations of a state-of-the-art unsupervised multi-object segmentation approach. We argue that text line recognition should be used as a benchmark for this task in future work.

Because unsupervised performances are not completely satisfactory, we combine this approach with a weak supervision from line-level transcriptions. Transcriptions are widely available and easy to produce with dedicated software, e.g. [24], and we show this dramatically improves the results, while preserving their interpretability. We believe that similar weak (i.e., image-level) annotations could also be considered for other images decomposition problems.

Contributions. To summarize, we present:

- a deep generative approach to text line analysis, inspired by deep unsupervised multi-object segmentation approaches and adapted to work in both a weakly supervised and unsupervised setting,
- a demonstration of the potential of our approach in challenging applications, particularly ciphered documents and paleographic analysis,
- experiments on three very different datasets: a printed volume of the Google1000 dataset [45, 19], the Copiale cipher [2, 27] and historical handwritten charters from the 12th and early 13th century [6, 43].

Our complete implementation can be found at github.com/ysig/learnable-typewriter.

2. Related Work

Text recognition. Image Text Recognition, including Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR), is a classic pattern recognition problem, and one of the earliest successful application of deep learning [31, 32]. The mainstream approaches for text line recognition rely on discriminative supervised learning. Typically, a Convolutional Neural Network (CNN) encoder will map the input image to a sequence of features and a decoder will associate them to the ground truth, e.g. through a recurrent architecture trained with a Connectionist Temporal Classification (CTC) loss [15, 16, 39, 4, 8], or a transformer trained with cross entropy [25, 33].

More related to our work, ScrabbleGAN [12] proposed a generative adversarial approach for semi-supervised text recognition, but their method is neither able to reconstruct an input text line nor to decompose it into individual characters. Also related are approaches which use already annotated sprites (referred to as exemplars or supports) to perform OCR/HTR [49, 42] by matching them to text lines. Recent unsupervised approaches, either cluster input images in a feature space [2] or rely on an existing text corpus of the recognized language [19].

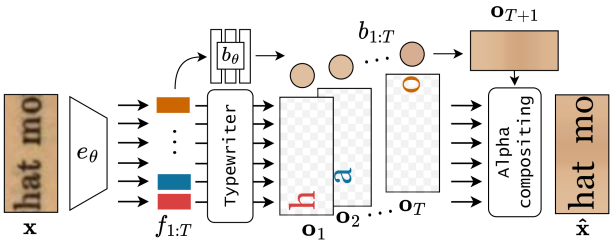
Closest to our work are classical prototype-based methods [28, 29, 46, 1] and in particular the Ocular system [3] which follows a generative probabilistic approach to jointly model text and character fonts in binarized documents, and is optimized through Expectation Maximization (EM). Different from us, it also relies on a pre-trained n-gram language model, originally from the english language and later extended to multiple languages [13]. Other approaches rely on language models to identify characters [30, 3, 19]. However, language models do not exist for unknown ciphers, or historical manuscripts which are often strongly abbreviated. Instead, we propose to disambiguate sprites by relying on line level transcriptions.

Unsupervised multi-object segmentation. Unsupervised multi-object segmentation refers to a family of approaches that decompose and segment scenes into multiple objects in an unsupervised manner [26]. Some techniques perform decomposition by computing pixel level segmentation masks over the whole input image [5, 10, 18, 17, 47], while others focus on smaller regions of the input image and learn to compose objects in an iterative fashion, mostly using a recurrent architecture [7, 9, 11, 23]. All of these techniques can isolate objects by producing segmentation masks, but our goal is also to capture recurring visual elements.

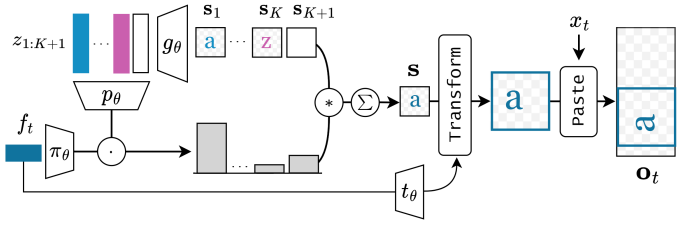
We thus build on techniques that explicitly model the objects located inside the input image, by associating them to a set of image prototypes referred to as sprites [37, 41]. Sprites are color images with an additional transparency channel and are associated to transformation prediction networks that are used to compose them onto a target canvas. However, DTI-Sprites [37] can only predict a small amount of sprites for a collection of fixed-size images and fails to scale when the number of objects inside each image scales to those of real documents. At the same time, MarioNette [41] suffers from a high reconstruction error and fuzzy sprites that sub-optimally reconstruct a toy text dataset.

3. The Learnable Typewriter

Given a collection of text lines written using consistent font or handwriting, our goal is to learn the shape of all the



(a) Overview of our full pipeline



(b) Details of our Typewriter module

Figure 2: **Overview. (a)** An image is encoded into a sequence of features, each decoded by the Typewriter module into image layers. They are then fused by alpha compositing with a predicted uniform background. **(b)** The Typewriter module takes a feature as input, computes sprites and associated probabilities from learned latent codes, and composes them into a composite sprite that is transformed and positioned onto an image-sized canvas.

characters it contains and a deep network that predicts the exact way these characters were used to generate any input text line. Since complete supervision (*i.e.*, the position and shape of every character used in a document) for such a task would be extremely costly to obtain, we propose to proceed in an analysis-by-synthesis fashion and to build on sprite-based unsupervised image decomposition approaches [41, 37] which jointly learn a set of character images - called *sprites* - and a network that transforms and positions them on a canvas in order to reconstruct input lines. Due to the potential ambiguity in the definition of sprites, we introduce a complementary weak-supervision from line-level transcriptions.

In this section, we first present an overview of our image model and approach (Section 3.1). Then, we describe the deep architecture we use (Section 3.2). Finally, we discuss our loss and training procedure (Section 3.3).

Notations. We write $a_{1:n}$ the sequence $\{a_1, \ldots, a_n\}$, and use bold letters $\mathbf{a}$ for images. An RGBA image $\mathbf{a}$ corresponds to an RGB image denoted by $\mathbf{a}^c$, alongside an alpha-transparency channel denoted by $\mathbf{a}^\alpha$. We use θ as a generic notation for network parameters and thus any character indexed by θ, e.g., a_θ, is a network.

3.1. Overview and image model

Figure 2a presents an overview of our pipeline. An input image $\mathbf{x}$ of size $H \times W$ is fed to an encoder network e_θ generating a sequence of T features $f_{1:T}$ associated to uniformly-spaced locations $x_{1:T}$ in the image. Each feature f_t is processed independently by our *Typewriter* module (Section 3.2) which outputs an RGBA image $\mathbf{o}_t$ corresponding to a character. The images $\mathbf{o}_{1:T}$ are then composited with a canvas image we call $\mathbf{o}_{T+1}$. This canvas image $\mathbf{o}_{T+1}$ is a completely opaque image (zero transparency). Its colors are predicted by a Multi-Layer Perceptron (MLP) b_θ which takes as input the features $f_{1:T}$ and outputs RGB values $b_{1:T}$. All resulting images $\mathbf{o}_{1:T+1}$ can be seen as ordered image layers and are merged using alpha compositing, as proposed by both [37, 41]. More formally, the reconstructed image $\hat{\mathbf{x}}$ can be written:

$$\hat{\mathbf{x}} = \sum_{t=1}^{T+1} \Big[\prod_{j<t} (1 - \mathbf{o}_j^\alpha) \Big] \mathbf{o}_t^\alpha \mathbf{o}_t^c. \tag{1}$$

In practice, we randomize the order of $\mathbf{o}_{1:T}$ in the compositing operation to reduce overfitting, as advocated by the MarioNette approach [41]. The full system is differentiable and can be trained end-to-end.

3.2. Typewriter Module

We now describe in detail the Typewriter module, which takes as input a feature f from the encoder and its position x, and outputs an image layer $\mathbf{o}$, to be composited. An overview of the module is presented in Figure 2b. On a high level, it is similar to the MarioNette architecture [41], but handles blanks (*i.e.*, the generation of a completely transparent image) in a different way and has a more flexible deformation model, similar to the one used in DTI-Sprites [37]. More specifically, the module learns jointly RGBA images called *sprites* corresponding to character images, and networks that use the features f to predict a probability for each sprite and a transformation of the sprite. We detail how we obtain the following three elements: the set of K parameterized sprites, the sprites compositing and the transformation model.

Sprite parametrization. We model characters as a set of K sprites which are defined using a generator network. More specifically, we learn K latent codes $z_{1:K}$ which are used as an input to a generator network g_θ in order to generate sprites $\mathbf{s}_{1:K} = g_\theta(z_{1:K})$. These sprites are images with a single channel that corresponds to their opacity. Similar to DTI-Sprites [37], we model a variable number of sprites with an empty (*i.e.*, completely transparent) sprite which we write $\mathbf{s}_{K+1}$. In comparison with directly learning sprites in the image space as in DTI-Sprites [37], we found that using a generator network yields faster and better convergence.

Sprite probabilities and compositing. To predict a probability p_k for each sprite $\mathbf{s}_k$, each latent code z_k is associated through a network p_θ to a probability feature $z_k^p = p_\theta(z_k)$ of the same dimension D as the encoder features ($D = 64$ in our experiments). We additionally optimize directly a probability feature z_{K+1}^p which we associate to the empty sprite. Given a feature f predicted by the encoder, we predict the probability p_k of each sprite $\mathbf{s}_k$ by computing the dot product between the probability features $z_{1:K+1}^p$ and a learned projection of the feature $\pi_\theta(f)$, and applying a $\mathrm{softmax}$ to the result:

$$p_{1:K+1}(f) = \mathrm{softmax}\left(\lambda z_{1:K+1}^p \cdot \pi_\theta(f)^T\right), \qquad (2)$$

where $\cdot$ is the dot product applied to each element of the sequence, $\lambda = 1/\sqrt{D}$ is a scalar temperature hyper-parameter, and the softmax is applied to the resulting vector. We use these probabilities to combine the sprites into the weighted average $\mathbf{s} = \sum_{k=1}^{K} p_k g_\theta(z_k)$. Note that this compositing can be interpreted as attention operation [44]:

$$\mathbf{s} = \mathrm{attention}(\bar{Q}, \bar{K}, \bar{V}) = \mathrm{softmax}\left(\frac{\bar{Q}\bar{K}^T}{\sqrt{D}}\right)\bar{V}, \qquad (3)$$

with $\bar{Q} = \pi_\theta(f)$, $\bar{K} = p_\theta(z_{1:K+1})$, $\bar{V} = g_\theta(z_{1:K+1})$, D the dimension of the features, and by convention $g_\theta(z_{K+1})$ is the empty sprite and $p_\theta(z_{K+1}) = z_{K+1}^p$.

We actually show that directly optimizing $z_{1:K}^p$ instead of learning to predict the probability features $z_{1:K}^p$ from the sprite latent codes $z_{1:K}$, similar to MarioNette [41], yields similar results. Note that we learn a probability code z_{K+1}^p to compute the probability of empty sprites instead of having a separate mechanism as in MarioNette [41] because it is critical for our supervised loss (see Sec. 3.3).

Positioning and coloring. The final step of our module is to position the selected sprite in a canvas of size $H \times W$ and to adapt its color. We implement this operation as a sequence of a spatial transformer [22] and a color transformation, similar to DTI-Sprites [37]. More specifically, the feature f is given as input to a network t_θ that predicts three parameters for the color of the sprite and three parameters for isotropic scaling and 2D-translation that are used by a spatial transformer [22] to deform $\mathbf{s}$. Finally, using the location x associated to the feature f, we paste the deformed colored sprite onto a background canvas of size $H \times W$ at position x to obtain a reconstructed RGBA image layer $\mathbf{o}$. Positioning the sprites with respect to the position of the associated local features helps us obtain results co-variant to translations of the text lines and independent of the line size. To produce the background canvas, the features $f_{1:T}$ are first each passed through a shared MLP b_θ, to produce background colors $b_{1:T}$. We then use bi-linear interpolation to upscale these T colors to fit the size of the input image. Details on the parametrization of the transformation networks are presented in the supplementary material.

3.3. Losses and training details

Our system is designed in an analysis-by-synthesis spirit, and thus relies mainly on a reconstruction loss. This reconstruction loss can be complemented by a loss on the selected sprites in the supervised setting where each text line is paired with a transcription. In the following, we define these losses for a single text line image and its transcription, using the notations of the previous section.

Reconstruction loss. Our core loss is a simple mean square error between the input image $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$ predicted by our system as described in Sec. 3.1:

$$\mathcal{L}_{\text{rec}}(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{x} - \hat{\mathbf{x}}\|^2. \qquad (4)$$

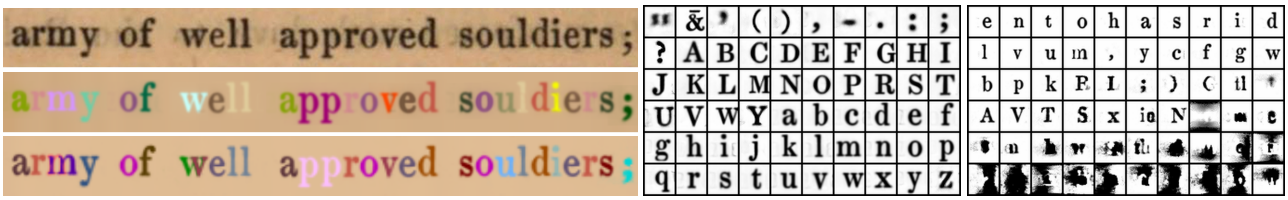
In the unsupervised setting, we use this loss alone without any additional regularization.

Weakly supervised loss. The intrinsic ambiguity of the sprite decomposition problem may result in sprites that do not correspond to individual characters. Using line-level annotation is an easy way to resolve this ambiguity. We find that simply adding the classical CTC loss [15] computed on the sprite probabilities to our reconstruction loss is enough to learn sprites that exactly correspond to characters. More specifically, we chose the number of sprites as the number of different characters and associate arbitrarily each sprite to a character and the empty sprite to the separator token of the CTC. Then given the one-hot line-level annotation y and the predicted sprite probabilities $\hat{y} = (p_{1:K+1}(f_1), ..., p_{1:K+1}(f_T))$, we optimize our system's parameters by minimizing:

$$\mathcal{L}_{\text{sup}}(\mathbf{x}, y, \hat{\mathbf{x}}, \hat{y}) = \mathcal{L}_{\text{rec}}(\mathbf{x}, \hat{\mathbf{x}}) + \lambda_{\text{ctc}}\mathcal{L}_{\text{ctc}}(y, \hat{y}) \qquad (5)$$

where λ_{ctc} is a hyper-parameter and $\mathcal{L}_{\text{ctc}}(y, \hat{y})$ is the CTC loss computed between the ground-truth y and the predicted probabilities $\hat{y}$. In our experiments we have used $\lambda_{\text{ctc}} = 0.1$ for printed text and $\lambda_{\text{ctc}} = 0.01$ for handwritten text.

Implementation and training details. We train on the Google1000 [45] and Fontenay [43] datasets with lines of height $H = 64$ and on the Copiale dataset [27] with $H = 92$. The generated sprites $\mathbf{s}_{1:K}$ are of size $H/2 \times H/2$. In the supervised setting, we use as many sprites as there are characters, and in the unsupervised we set $K = 60$ for Google1000 and $K = 120$ for the Copiale cipher. In the supervised case we train for 100 epochs on Google1000 and for 500 epochs on Copiale with a batch size of 16, and we select the model that performs best on the validation set for



(a) Input, supervised and unsupervised semantic segmenation. (b) Supervised Sprites. (c) Unsupervised Sprites.

Figure 3: **Results on a printed document (Google1000).** In both the supervised and unsupervised setting our method produces meaningful sprites and accurate reconstructions (3a). We show the 60 most used sprites in alphabetic ordering in the supervised setting (3b) and ordered by frequency in the unsupervised one (3c). See text for details and the supplementary material for more reconstructions.

evaluation. In the unsupervised setting we use line crops of width $W = 2H$ and train for 1000 epochs on Google1000 and for 5000 on the Copiale cipher with a batch size of 32 and use the final model. The number of epochs is much higher in the unsupervised case than in the supervised case because the network sees only a small crop of each line at each epoch, but each epoch is much faster to perform.

Our encoder network is a ResNet-32-CIFAR10 [20], that is truncated after layer 3 with a Gaussian feature pooling described in supplementary material. For our unsupervised experiments, we use as generator g_θ the U-Net architecture of Deformable Sprites [48] which converged quickly, and for our supervised experiments a 2-layer MLP similar to MarioNette [41] which produces sprites of higher quality. The networks π_θ and p_θ are a single linear layers followed by layer-normalization. We use the AdamW [34] optimizer with a learning rate of 10^{-4} and apply a weight-decay of 10^{-6} to the encoder parameters. At inference we select the sprites with the highest probabilities instead of using a softmax.

4. Experiments

4.1. Datasets and metrics

Datasets. We experiment with three datasets with different characteristics: Google1000 [45], the Copiale cipher [27] and Fontenay manuscripts [43, 6]:

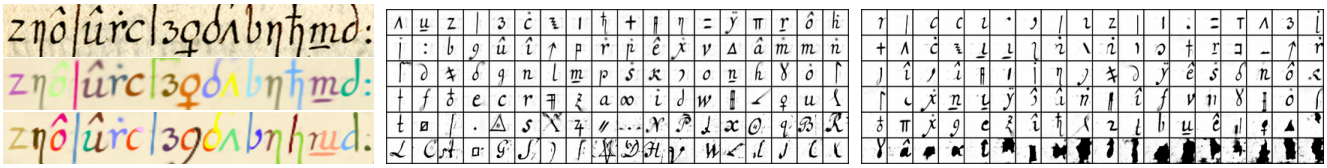
• *Google1000.* The Google1000 dataset contains scanned historical printed books, arranged into Volumes [45]. We use the English Volume 0002 which we process with the preprocessing code of [19], using 317 out of 374 pages and train-val-test set with 5097-567-630 lines respectively. This leads to a total number of 83 distinct annotated characters. Although supervised printed font recognition is largely considered a solved problem, and the annotation for this dataset are actually the result of OCR, this document is still challenging for an analysis-by-synthesis approach, containing artifacts such as ink bleed, age degradation, as well as variance in illumination and geometric deformations.

• *Copiale cipher.* The Copiale cipher is an oculist German text dating back to a 18th century secret society [27]. Opposite to Baro et al. [2] which uses a binarized version of the dataset, we train our model on the original text-line images, which we segmented using docExtractor [35] and manually assigned to the annotations, respecting the train-val-test split of Baro et al. [2] with 711-156-908 lines each. The total number of distinct annotated characters is 112. This dataset is more challenging than printed text because because it is handwritten, which introduces some variability in the character shapes, and because of the large number of characters.

• *Fontenay manuscripts.* The Fontenay dataset contains digitized charters that originate from the Cistercian abbey of Fontenay in Burgundy (France) [43, 6] and were created during the 12th and early 13th century. Each document has been digitized and each line has been manually segmented and transcribed. For our experiments, we selected a subset of 14 different documents sharing a similar script which falls into the family of praegothica scripts. These correspond to 163 lines, using 47 distinct characters. While they were carefully written and preserved, these documents are still very challenging (Figure 6). They exhibit degradation, clear intra-document letter shape variations, and letters can overlap or be joined by ligature marks. Moreover, each document represents only a small amount of data, e.g., the ones used in Figure 6 contain between 8 and 25 lines.

Metrics. Our goal is to capture the shape of all characters and position them precisely in each text line. Such fine-grained annotation is however not available in existing datasets. Instead, to provide a quantitative evaluation of our models, we report L2 reconstruction error ('Rec.' in the tables) and Character Error Rate (CER). CER is the standard metric for Optical Character Recognition (OCR). Given ground-truth and predicted sequences of characters, σ and $\hat{\sigma}$, it is defined as the minimum number of substitutions S, deletions D, and insertions I of characters necessary to match the predicted sequence $\hat{\sigma}$ to the ground truth sequence σ, normalized by the size of the ground truth sequence $|\sigma|$:

$$CER(\sigma, \hat{\sigma}) = \frac{S + D + I}{|\sigma|}. \quad (6)$$



(a) Input, supervised and unsupervised. (b) Supervised Sprites. (c) Unsupervised Sprites.

Figure 4: **Results on the Copiale cipher [27].** Despite the high number of characters and their variability, our method learns meaningful sprites and performs accurate reconstructions in both settings (4a). We show the 108 most used sprites sorted by frequency in the supervised (4b) and the unsupervised (4c) settings.

For simplicity, we ignore spaces. Predictions are obtained by selecting at every position the character associated to the most likely sprite. In the supervised setting, the association between sprites and characters is fixed at the beginning of training. In the unsupervised setting, we associate every sprite to a single character using a simple assignment strategy described in supplementary material. More complex assignments, for example associating sprite bi-grams to individual characters, or even incorporating their relative positions, could be considered for a recognition performance boost. However, since OCR is not our main goal but simply a proxy measure, this falls out of the scope of our work.

4.2. Qualitative results

Examples of semantic segmentation and sprites in the supervised and unsupervised setting on Google1000 and Copiale are shown in Figures 3 and 4 respectively.

In the unsupervised setting, several sprites (Figures 3c and 4c) can be used to reconstruct a single character. For example, the 'n' and 'm' sprites are joined in order to better reconstruct 'm' in Google1000. To account for appearance variation, multiple sprites are learned to reconstruct the most frequent character, e.g. 'e' for Google and 'c' in Copiale. This effects are even stronger in the handwritten Copiale dataset, where generic sub-character strokes are learned and used together to better model characters' variations.In both datasets, a portion of the least used sprites are not well optimized, do not correspond to characters, and are not used by the network. These behaviors are expected in a completely unsupervised setting, because of the highly unbalanced statistics of the character frequencies and ambiguity of the reconstruction: without additional supervision, there is a clear benefit for the network to model well variations of common characters, and to approximate or discard rare ones. This is a core limitation of existing unsupervised image decomposition approaches, and a motivation for the introduction of our weakly supervised setting.

In the (weakly) supervised setting, the sprites (Figures 3b and 4b) closely correspond to the characters, with the exception of very rare characters like the capital 'Z' character for Google1000 (as can be seen in supplementary material), while reconstruction is of very high quality and each character is reconstructed with the expected sprite.

4.3. Quantitative results

Our quantitative results and ablations for Google1000 and Copiale are reported in Tables 1 and 2 respectively.

For Google1000, the CER in the supervised setting is close to perfect, while it is 7.7% for the unsupervised setting. To provide baselines for these performances, we trained on our data (i) ScrabbleGAN [12], a supervised method with a standard recognizer and an additional generator module, (ii) FontAdaptor [49], a recent 1-shot method that learns to match single character examplars to text lines, and (iii) an adaptation of the unsupervised DTI-Sprites [36] to text lines which we detail in the supplementary material (we also show in the supplementary material that vanilla MarioNette [41] provides clearly worse results.). Our unsupervised approach performs clearly better than our adaptation of DTI-Sprites and is almost on par with the 1-shot FontAdaptor, while our weakly supervised approach is almost on par with ScrabbleGAN. Our adaptation of DTI-Sprites is better reconstruct images, but the learned sprites are much less meaningful, as shown by the poor CER performance. Interestingly, reconstruction is much better when using supervision, which hints that a better optimization scheme might help improving unsupervised performances. We also evaluated the effect of varying the number of sprites K in the unsupervised setting. For K smaller than the actual number of characters (83), namely $K = 21$ and $K = 41$, we have a significant performance drop of 10% and 26% CER respectively, while increasing the number of characters to 166 and 332 doesn't significantly boost performances.

On the Copiale dataset, we compare our results with HTRbyMatching [42], a few-shot approach developed specifically for cipher recognition, using the same train/val/test splits. HTRbyMatching was evaluated on a wide range of few-shot scenarios, ranging from a scenario similar to FontAdaptor where a single exemplar is available for every character, to one where 5 exemplars are available for each character together with 5 completely annotated pages. Reported results are only for confident character predictions with different

Method	Type	Rec. $\times 10^3$	CER
DTI-Sprites [37]	unsup.	2.54	18.4 %
FontAdaptor [49]	1-shot	-	6.7 %
ScrabbleGAN [12]	sup.	-	0.6 %
Learnable Typewriter	sup.	3.5 ± 0.1	$0.85 \pm 0.03\%$
w\o shared z_k	sup.	3.3 ± 0.1	$0.89 \pm 0.06\%$
w\o p_θ	sup.	3.5 ± 0.1	$0.99 \pm 0.05\%$
w\o g_θ	sup.	3.4 ± 0.1	$0.88 \pm 0.04\%$
Learnable Typewriter	unsup.	7.1 ± 0.4	$7.7 \pm 0.6\%$
w\o shared z_k	unsup.	7.4 ± 0.4	$8.0 \pm 0.2\%$
w\o p_θ	unsup.	7.0 ± 0.3	$7.7 \pm 2.0\%$
w\o g_θ	unsup.	10.5 ± 0.7	$27.0 \pm 2.2\%$

Table 1: **Quantitative results and ablation on Google1000 [45].** We report CER and L2 reconstruction error for different approaches. For our method, we report average over 5 runs and standard deviation.

Method	Type	Rec. $\times 10^2$	CER
HTRbyMatching [42]	few-shot	-	$10 - 47\%^*$
Learnable Typewriter	sup.	1.81 ± 0.01	$4.2 \pm 0.3\%$
w\o shared z_k	sup.	1.79 ± 0.01	$4.0 \pm 0.1\%$
w\o p_θ	sup.	1.77 ± 0.02	$4.7 \pm 0.1\%$
w\o g_θ	sup.	1.96 ± 0.07	$4.2 \pm 0.2\%$
Learnable Typewriter	unsup.	1.93 ± 0.02	$52.6 \pm 1.7\%$
w\o shared z_k	unsup.	1.89 ± 0.02	$47.6 \pm 2.8\%$
w\o p_θ	unsup.	1.81 ± 0.06	$51.9 \pm 2.0\%$
w\o g_θ	unsup.	3.99 ± 0.14	$80.6 \pm 0.9\%$

Table 2: **Quantitative results on Copiale [27].** We report CER and reconstruction error for a baselines and our method. For our method, we report average over 5 runs and standard deviation. *See text for details.

confidence thresholds, but summing the error rate of the predicted symbols and the percentage of non-annotated symbols, one can estimate the CER to vary between 10% and 47% depending on the scenario. This is consistent with the quantitative results we obtain with our approach, which are much better in the supervised setting (4.2%) and worse in the completely unsupervised one (52.6%). The low performance of the unsupervised approach is consistent with the qualitative results: given that many characters are reconstructed are reconstructed by sub-character sprites, one would have to associate sprite bi-grams to characters in order to obtain good CER performances. Interestingly, the reconstruction error is similar in the supervised and unsupervised setting, hinting that for this specific dataset, optimizing the reconstruction quality might not be enough to obtain relevant decomposition without additional priors. These results enables to quantify and analyze a limitation of unsupervised image decomposition approaches on a more challenging dataset.

Note that the goal of our approach is not to boost CER performances - which in any case would be meaningless on Google1000 where the ground truth is already the result of an OCR model - but instead to learn character models and image decomposition, and all these comparisons should be considered as sanity checks. Designing post-processing to improve CER is possible, for example we tested a simple post-processing associating new sprites to the most frequent bi-grams and tri-grams, that leads to an improved CER for Copiale of 29.9%. However, we think it is more interesting to see this metric as a tool to evaluate the raw output of unsupervised decomposition models.

In particular, we performed on both datasets an ablation of the architecture to better understand which design choices are critical. Interestingly, our results show that both in the supervised and the unsupervised setting, not sharing the latent codes z_k between the generation network and the sprite selection and even completely removing the probability network p_θ has limited influence on the performance, and these design choices of MarioNette [41] are not critical. Conversely, removing g_θ and directly learning prototypes as network parameters similar to DTI-Sprites [37] has little impact in the supervised case but leads to significant drops in performance. A more detailed analysis of training curves reveals that training is slower and overfits. While it might be possible to fix this issue by adapting the learning scheme for the prototypes, it shows it is easier to learn the prototypes through a generator network than optimizing them directly.

4.4. Application to paleography

To test our approach in a more challenging case and demonstrate its potential for paleographic analysis, we applied it on a collection of 14 historical charters from the Fontenay abbey [43, 6]. While they all use similar scripts from the Praegothica type, they also exhibit clear variations. One of the goals of a paleographic analysis would be to identify and characterize these variations. We focus on the variations in the shape of letters, which are quite challenging to describe with natural language. One solution would be to choose a specific example for each letter in each document or to have a paleographer manually draw of a 'typical' one. However, this is very time consuming and might reflect priors or bias from the paleographer in addition to the actual variations. Instead, we propose to fine-tune our Learnable Typewriter approach on each document and visualize the sprites associated to each character and each document. Because of the difficulty of the dataset, we focus on the results of our supervised setting.

Figure 5 visualizes the sprites obtained for five different documents from the characters 'a' to 'h' and Figure 6 highlights different aspects of the results. Figure 6a emphasizes the fact that the differences in the learned sprites correspond

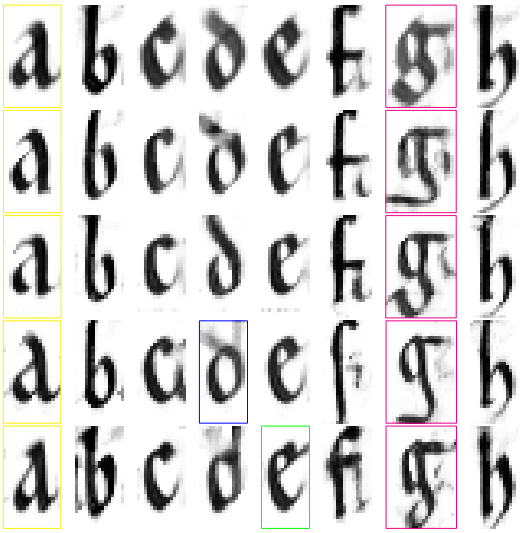
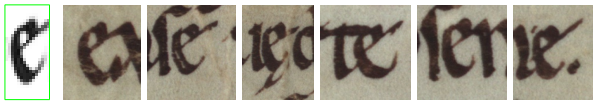
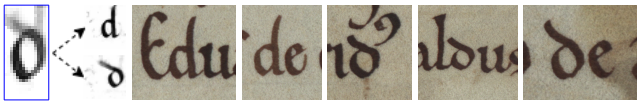


Figure 5: **Sprites learned for similar documents in Praegothica script.** Each line corresponds to a different document. Looking at any column, one can notice the small differences that characterise the handwriting in each document. Colored boxed correspond to cases analysed in more details in Figure 6.

(a) 'a' and 'g' sprite for each document and associated example of the character. Note how the variations of the descending part of the 'g' sprites closely match the variations observed in the documents. Also note the subtle variations of the 'a' which are clear in the sprites but would be hard to notice and describe from the original images for a non-expert.



(b) The appearance variations of individual instances associated to the 'e' character in the document are accurately visually summarized by the sprite.



(c) The double appearance of the ascending line of the 'd' sprite shown on the left is related to the co-existence of two different kinds of 'd' in the document, as shown in the examples on the right. We can actually learn both appearances of 'd', shown after the arrows, if we model every character using two sprites.

Figure 6: The sprites summarize the key attributes of a character in each specific document, averaging its variations. Note the complexity of the documents: characters can overlap or be connected ligature, the parchment is often stained, and there are important intra-document character variations.

to actual variations in the different documents, whether subtle, such as for the 'a' sprite, or clearer, such as for the descending part of the 'g' sprite. Figure 6b shows how a sharp sprite can be learned for the character 'e', summarizing accurately its shape despite variations in the different occurrences. Finally, Figure 6c shows the case of a document in which two types of 'd' co-exist. In this case, the learned sprite, shown on the left, reassembles an average of the two, with both versions of the ascending parts visible with intermediate transparency. Such a limitation could be overcome by learning several sprites per character. We thus experimented with learning two per character, simply by summing their probabilities when optimizing the CTC-loss. We find that when different appearances of the same letter exist, the two sprites learn two different appearances, and we show the example of the two different learned 'd' sprites on the right of the original one.

Our approach could benefit paleographic analysis in more ways than simply analyzing the characters shapes. Indeed, our model also gives access to the position and scale variation for each letter. This would enable a quantitative analysis of more global appearance factors of the text, related to the space between letters or their respective size variations. Because they would be tremendously tedious to annotate, such variations have rarely been quantified, and their analysis could open new research topics, for example the study of the handwriting evolution of a single writer copying a book across several months. Another natural application of our approach is font or writer classification, which could be achieved either using a single model to compare errors statistics for the different letters or relative positions of bi-grams, or by training different models for different fonts or writers. The main advantage compared to most existing approaches would be the high interpretability of the predictions, which a user could easily validate.

5. Conclusion

We have presented a document-specific generative approach to document analysis. Inspired by deep unsupervised multi-object segmentation methods, we extended them to accurately model standard printed documents as well as much more complex ones, such as a handwritten ciphered manuscript or ancient charters. We outlined that a completely unsupervised approach suffers from the ambiguity of the decomposition problem and imbalance characters distributions. We thus extended these approaches using weak supervision to obtain high-quality results. Finally, we demonstrated the potential of our Learnable Typewriter approach for a novel application: paleographic analysis.

Acknowledgments

We would like to thank Malamatenia Vlachou and Dominique Stutzmann for sharing ideas, insights and data for applying our method in paleography; Vickie Ye and Dmitriy Smirnov for useful insights and discussions; Romain Loiseau, Mathis Petrovich, Elliot Vincent, Sonat Baltacı for manuscript feedback and constructive insights. This work was partly supported by the European Research Council (ERC project DISCOVER, number 101076028), ANR project EnHerit ANR-17-CE23-0008, ANR project VHS ANR-21-CE38-0008 and HPC resources from GENCI-IDRIS (2022-AD011012780R1, AD011012905).

References

[1] Henry S Baird. Model-directed document image analysis. In *Proceedings of the Symposium on Document Image Understanding Technology*, volume 1, page 3, 1999. 1, 2

[2] Arnau Baró, Jialuo Chen, Alicia Fornés, and Beáta Megyesi. Towards a Generic Unsupervised Method for Transcription of Encoded Manuscripts. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, pages 73–78, Brussels Belgium, May 2019. ACM. 1, 2, 5

[3] Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. Unsupervised Transcription of Historical Documents. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 207–217, 2013. 1, 2

[4] Théodore Bluche and Ronaldo Messina. Gated convolutional recurrent neural networks for multilingual handwriting recognition. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 646–651. IEEE, 2017. 2

[5] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019. 1, 2

[6] Jean-Baptiste Camps, Chahan Vidal-Gorène, Dominique Stutzmann, Marguerite Vernet, and Ariane Pinche. Data Diversity in handwritten text recognition: challenge or opportunity? In DH2022 Local Organizing Committee, editor, *Digital Humanities 2022. Conference Abstracts (The University of Tokyo, Japan, 25-29 July 2022)*, pages 160–165. Tokyo, 2022. 1, 2, 5, 7

[7] Eric Crawford and Joelle Pineau. Spatially invariant unsupervised object detection with convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3412–3420, 2019. 1, 2

[8] Arthur Flor de Sousa Neto, Byron Leite Dantas Bezerra, Alejandro Héctor Toselli, and Estanislau Baptista Lima. Htr-flor: a deep learning system for offline handwritten text recognition. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 54–61. IEEE, 2020. 2

[9] Fei Deng, Zhuo Zhi, Donghun Lee, and Sungjin Ahn. Generative scene graph networks. In *International Conference on Learning Representations*, 2020. 1, 2

[10] Patrick Emami, Pan He, Sanjay Ranka, and Anand Rangarajan. Efficient iterative amortized inference for learning symmetric and disentangled multi-object representations. In *International Conference on Machine Learning*, pages 2970–2981. PMLR, 2021. 1, 2

[11] S. M. Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E. Hinton. Attend, Infer, Repeat: Fast Scene Understanding with Generative Models. In *Advances in Neural Information Processing Systems*, volume 29, Aug. 2016. 1, 2

[12] Sharon Fogel, Hadar Averbuch-Elor, Sarel Cohen, Shai Mazor, and Roee Litman. ScrabbleGAN: Semi-Supervised Varying Length Handwritten Text Generation. *arXiv:2003.10557 [cs]*, Mar. 2020. 2, 6, 7

[13] Dan Garrette, Hannah Alpert-Abrams, Taylor Berg-Kirkpatrick, and Dan Klein. Unsupervised Code-Switching for Multilingual Historical Document Transcription. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1036–1041, Denver, Colorado, 2015. Association for Computational Linguistics. 2

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[15] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA, 2006. Association for Computing Machinery. 2, 4

[16] Alex Graves and Jürgen Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. *Advances in neural information processing systems*, 21:545–552, 2008. 1, 2

[17] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *International Conference on Machine Learning*, pages 2424–2433. PMLR, 2019. 1, 2

[18] Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. Neural expectation maximization. *Advances in Neural Information Processing Systems*, 30, 2017. 1, 2

[19] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Learning to Read by Spelling: Towards Unsupervised Text Recognition. *arXiv:1809.08675 [cs]*, Dec. 2018. 1, 2, 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[21] Judith Hochberg, Patrick Kelly, Timothy Thomas, and Lila Kerns. Automatic script identification from document images using cluster-based templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):176–181, 1997. 1

[22] Max Jaderberg, Karen Simonyan, and Andrew Zisserman. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. 4

[23] Jindong Jiang and Sungjin Ahn. Generative neurosymbolic machines. *Advances in Neural Information Processing Systems*, 33:12572–12582, 2020. 1, 2

[24] Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 4, pages 19–24. IEEE, 2017. 2

[25] Lei Kang, Pau Riba, Marçal Rusiñol, Alicia Fornés, and Mauricio Villegas. Pay attention to what you read: Non-recurrent handwritten text-line recognition. *arXiv preprint arXiv:2005.13044*, 2020. 2

[26] Laurynas Karazija, Iro Laina, and Christian Rupprecht. ClevrTex: A Texture-Rich Benchmark for Unsupervised Multi-Object Segmentation. *arXiv:2111.10265 [cs]*, Nov. 2021. 2

[27] Kevin Knight, Beata Megyesi, and Christiane Schaefer. The Copiale Cipher. In *Proceedings of the ACL Workshop on Building and Using Comparable Corpora*, pages 2–9, 2011. 1, 2, 4, 5, 6, 7

[28] Gary E Kopec and Mauricio Lomelin. Document-specific character template estimation. In *Document Recognition III*, volume 2660, pages 14–26. SPIE, 1996. 1, 2

[29] Gary E Kopec and Mauricio Lomelin. Supervised template estimation for document image decoding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(12):1313–1324, 1997. 1, 2

[30] Gary E Kopec, Maya R Said, and Kris Popat. N-gram language models for document image decoding. In *Document Recognition and Retrieval IX*, volume 4670, pages 191–202. SPIE, 2001. 2

[31] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 2

[32] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[33] Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv preprint arXiv:2109.10282*, 2021. 1, 2

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[35] Tom Monnier and Mathieu Aubry. docExtractor: An off-the-shelf historical document element extraction. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 91–96, Dortmund, Germany, Sept. 2020. IEEE. 5

[36] Tom Monnier, Thibault Groueix, and Mathieu Aubry. Deep Transformation-Invariant Clustering. In *NeurIPS*, Oct. 2020. 6

[37] Tom Monnier, Elliot Vincent, Jean Ponce, and Mathieu Aubry. Unsupervised Layered Image Decomposition into Object Prototypes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8640–8650, Apr. 2021. 1, 2, 3, 4, 7

[38] Joseph C Nolan and Robert Filippini. Method and apparatus for creating a high-fidelity glyph prototype from low-resolution glyph images, Apr. 20 2010. US Patent 7,702,182. 1

[39] Joan Puigcerver. Are multidimensional recurrent layers really necessary for handwritten text recognition? In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 67–72. IEEE, 2017. 2

[40] Pradyumna Reddy, Paul Guerrero, and Niloy J Mitra. Search for concepts: Discovering visual concepts using direct optimization. *arXiv preprint arXiv:2210.14808*, 2022. 1

[41] Dmitriy Smirnov, Michael Gharbi, Matthew Fisher, Vitor Guizilini, Alexei A. Efros, and Justin Solomon. MarioNette: Self-Supervised Sprite Learning. *arXiv:2104.14553 [cs]*, Apr. 2021. 1, 2, 3, 4, 5, 6, 7

[42] Mohamed Ali Souibgui, Alicia Fornés, Yousri Kessentini, and Crina Tudor. A few-shot learning approach for historical ciphered manuscript recognition. *CoRR*, abs/2009.12577, 2020. 2, 6, 7

[43] Dominique. Stutzmann. Fontenay dataset. original charters from fontenay before 1213 https://doi.org/10.5281/zenodo.6507963. 1, 2, 4, 5, 7

[44] 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. 4

[45] L. Vincent. Google Book Search: Document Understanding on a Massive Scale. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*, pages 819–823, Curitiba, Parana, Brazil, Sept. 2007. IEEE. 1, 2, 4, 5, 7

[46] Yihong Xu and George Nagy. Prototype extraction and adaptive ocr. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1280–1296, 1999. 1, 2

[47] Yanchao Yang, Yutong Chen, and Stefano Soatto. Learning to manipulate individual objects in an image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6558–6567, 2020. 1, 2

[48] Vickie Ye, Zhengqi Li, Richard Tucker, Angjoo Kanazawa, and Noah Snavely. Deformable sprites for unsupervised video decomposition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 5

[49] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Adaptive Text Recognition through Visual Matching. *ECCV 2020*, Sept. 2020. 2, 6, 7