Jointly Aligning Millions of Images with Deep Penalised Reconstruction Congealing

Roberto Annunziata
Onfido, UK
roberto.annunziata@onfido.com

Christos Sagonas
Onfido, UK
christos.sagonas@onfido.com

Jacques Cali*
Blue Prism
jqscali@gmail.com

Abstract

Extrapolating fine-grained pixel-level correspondences in a fully unsupervised manner from a large set of misaligned images can benefit several computer vision and graphics problems, e.g. co-segmentation, super-resolution, image edit propagation, structure-from-motion, and 3D reconstruction. Several joint image alignment and congealing techniques have been proposed to tackle this problem, but robustness to initialisation, ability to scale to large datasets, and alignment accuracy seem to hamper their wide applicability. To overcome these limitations, we propose an unsupervised joint alignment method leveraging a densely fused spatial transformer network to estimate the warping parameters for each image and a low-capacity auto-encoder whose reconstruction error is used as an auxiliary measure of joint alignment. Experimental results on digits from multiple versions of MNIST (i.e., original, perturbed, affNIST and infiMNIST) and faces from LFW, show that our approach is capable of aligning millions of images with high accuracy and robustness to different levels and types of perturbation. Moreover, qualitative and quantitative results suggest that the proposed method outperforms state-of-the-art approaches both in terms of alignment quality and robustness to initialisation.

Figure 1. Unsupervised joint alignment (a.k.a. *congealing*) results obtained by the proposed method on digit '2' from *affNIST* [43] and `Jennifer_Capriati` from LFW [18]. (a) input images before alignment (initialisation in red), (b) output images aligned with the proposed method.

1. Introduction

Establishing pixel-level correspondences between pair of images including instances of the same object category can benefit several important applications, such as motion estimation [33], medical imaging [1, 19], object recognition [13] and 3D reconstruction [20]. As a result, this has become a fundamental problem in computer vision [12, 14]. Typically, pixel-level correspondences between two images are computed by extracting sparse local feature descriptors (e.g., SIFT [29], HOG [11], SURF [5], SCIRD [2, 4]),

then matching the extracted descriptors, and finally pruning mismatches based on geometric constraints. Although this approach has been applied successfully in different domains, its performance can degrade significantly due to factors such as intra-class variations, non-rigid deformations, partial occlusions, illumination, image blur, and visual clutter. Recently, the representational power of Convolutional Neural Networks (CNNs) has been leveraged to improve the overall process. In particular, several CNN-based methods for learning powerful feature descriptors have been introduced [40, 47, 31]. More recently, end-to-end trainable CNNs for learning image descriptors as well as estimating the geometric transformation between the two images have been introduced in [36, 21]. The majority of previously pro-

---

*Contribution to this research project was entirely made while this co-author was at Onfido.

posed methods focus on the problem of finding pixel-level correspondences between a *pair* of images. However, a plethora of other tasks such as, co-segmentation, image edit propagation, video stabilisation and structure-from-motion, require global correspondences between a *set* of images containing a specific object. A straightforward way to address this problem is to identify pixel correspondences between each pair of images in the dataset and solve the problem in a sequential manner. However, this approach would be prone to important limitations, as (i) it would fail to take into account valuable cross-image appearance information (i.e. statistics of local patches across the entire dataset) during the optimisation; and (ii) the computational complexity of the problem would increase exponentially with the number of images, therefore significantly limiting the scalability to large datasets. Thus, estimating global correspondences of a set of images (image ensemble) by jointly aligning them in an unsupervised manner can be immensely valuable.

Congealing (joint image alignment) was originally introduced by Learned-Miller in [23]. His approach aligns (by estimating rigid transformations) an ensemble of images of a particular object by minimising the sum of entropies of pixel values at each pixel location. Although this method has been effectively applied to handwritten digits and magnetic resonance image volumes, it has shown some limitations, including slow and/or sometimes poor convergence and relatively high sensitivity to hyper-parameters. Later, Huang *et al.* improved the performance of [23] by using hand-crafted SIFT features [17]. To overcome the optimisation problems of the original congealing approach, Cox *et al.* [9, 10] proposed to utilise a reference image (i.e., template) and then minimise the sum of squared differences instead of the sum of entropies. This way, standard Gauss-Newton gradient descent method could be adopted to make the optimisation efficient. Later, motivated by lossy compression principles, Vedaldi *et al.* [44] proposed a joint alignment approach based on log-determinant estimation.

A common drawback of the aforementioned methods is that they cannot simultaneously handle variability in terms of illumination, gross pixel corruptions and/or partial occlusions. RASL, an image congealing method that overcomes this drawback was proposed in [35]. The key assumptions made in RASL and its multiple variants, e.g. [26, 8, 32] are that (i) an ensemble of well-aligned images of the same object is approximately low-rank and (ii) that gross pixel errors are sparsely distributed. Therefore, image congealing is performed by seeking a set of optimal transformations such that the ensemble of misaligned images is written as the superposition of two components i.e., a low-rank component and a sparse error component. RASL has been widely used for jointly aligning multiple images in different applications such as face landmarks localisation [37, 34, 39],
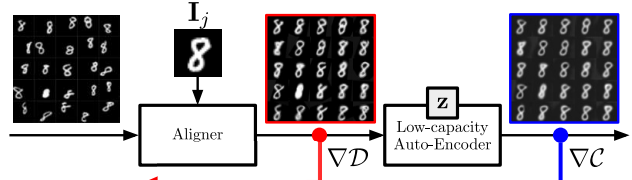

Figure 2. Block diagram of the proposed method. Black arrows correspond to forward pass, while red and blue to backpropagation.

pose-invariant face recognition [39, 38], and medical imaging [6]. Despite its wide applicability, it is worth noting that (i) RASL joint alignment performance can be severely affected by non-optimal initialisation and high intra-class variability in the image ensemble; (ii) scalability to large ensembles is limited by the formulation of the low-rank minimisation problem and related SVD-based sub-routines; and (iii) a new optimisation is required for every new image added to the ensemble. To address some of these limitations, t-GRASTA [15] and PSSV [32] have been recently proposed.

The first deep learning approach to unsupervised joint image alignment was proposed by Huang *et al.* [16]. A modified version of the convolutional restricted Boltzmann machine was introduced to obtain features that could better represent the image at differing resolutions, and that were specifically tuned to the statistics of the data being aligned. They then used those learnt features to optimise the standard entropy-based congealing loss and achieved excellent joint alignment results on the Labelled Faces in the Wild (LFW) benchmark.

Here, we propose a congealing method to solve large-scale joint alignment problems, which is of significant practical importance in light of the ever increasing availability of image data. The proposed method consists of two main modules: (i) the aligner and (ii) the low-capacity auto-encoder. Specifically, the joint alignment task is cast as a batch-based optimisation problem in which the aligner is used to estimate the global transformation required to warp each image to a reference. The alignment error is quantified via ℓ_1-norm between the transformed batch images and the reference. Motivated by the observation that a set of well-aligned images require less modelling capacity to be reconstructed well (e.g. reconstruction with low-rank bases [35]), the aligned batch is subsequently processed by a *low-capacity* auto-encoder and reconstruction errors are back-propagated to the aligner (a snapshot of the results is displayed in Fig. 1).

Contributions: In summary, the main contributions of this paper are: (i) a congealing method which is shown to be capable of handling large-scale joint alignment problems i.e., up to one million data points, simultaneously; (ii) a novel *differentiable* formulation of the congealing

problem, which combines the advantages of previously proposed similarity- and rank-based approaches and that can be easily optimised with Stochastic Gradient Descent (SGD), end-to-end; (iii) an extensive experimental evaluation of the proposed method and state-of-the-art approaches on several benchmark datasets, including digits and faces at different resolutions, assessing joint alignment performance and robustness to linear and non-linear geometric perturbations of different magnitude and type.

2. Methodology

In the following, we briefly summarise the approaches most related to ours. Then, we introduce the proposed method.

RASL. Let us assume we have N misaligned images $\{\mathbf{I}_i\}_{i=1}^N \in \mathbb{R}^{w \times h}$ of a particular object and let $\{\mathbf{p}_i\}_{i=1}^N$ be a set of transformations such that $\{\mathbf{I}_i^0 = \mathbf{I}_i \circ \mathbf{p}_i\}_{i=1}^N$ becomes a set of well-aligned images. If we define vec $: \mathbb{R}^{w \times h} \to \mathbb{R}^m$ as the operator that vectorises an image, the main assumption of the RASL method is that the matrix:

$$\mathbf{D} \circ \mathbf{P} = [\text{vec}(\mathbf{I}_1^0) \mid \cdots \mid \text{vec}(\mathbf{I}_N^0)] = \mathbf{A} \qquad (1)$$

will be approximately *low-rank*. However, in practice this assumption can be violated when the object of interest is affected by occlusion, shadows, and noise. Therefore, the authors assume that each aligned image is corrupted with non-Gaussian-but-sparse errors $\mathbf{E} \in \mathbb{R}^{m \times N}$, such that $\mathbf{D} \circ \mathbf{P} = \mathbf{A} + \mathbf{E}$. Given the observation of the misaligned and corrupted images, the goal is to estimate a set of transformations $\{\mathbf{p}_i\}_{i=1}^N$ such that the rank of the transformed noise-free images $\{\mathbf{I}_i \circ \mathbf{p}_i\}_{i=1}^N \in \mathbb{R}^{w \times h}$ becomes as small as possible. Formally,

$$\underset{\mathbf{A}, \mathbf{E}, \{\mathbf{p}_i\}_{i=1}^N}{\arg\min} \; \text{rank}(\mathbf{A}) \quad \text{s.t.} \; \mathbf{D} \circ \mathbf{P} = \mathbf{A} + \mathbf{E}, \; \|\mathbf{E}\|_0 \leq q,$$

$$(2)$$

where q controls the sparsity of the error matrix $\mathbf{E}$. Unfortunately, the non-convex and discontinuous nature of the optimisation problem in Eq. (2) makes it not directly tractable. To this end, an algorithm that provides a sub-optimal solution via iterative convex programming was proposed. As discussed in [35], this algorithm is limited by the following assumptions: (i) initial misalignment not too large, (ii) rank of the matrix $\mathbf{A}$ to be recovered not too high, and (iii) only a small fraction of all pixels affected by error. A further limitation is the scalability of the algorithm. In fact, the convex relaxation replacing the rank($\cdot$) with the nuclear norm requires a very expensive Singular Value Decomposition (SVD) computation at every optimisation step.

Least-Squares Congealing (LSC). This method [9, 10] has been specifically proposed for targeting large-scale joint alignment problems. Building on the success of Lucas-Kanade image alignment [30], the idea is to define a reference image $\mathbf{I}_j$ and align each of the remaining ones $\{\mathbf{I}_i\}_{i \neq j}$

to that reference. In general, this optimisation problem can be formulated as:

$$\underset{\mathbf{p}_{i \neq j}}{\arg\min} \sum_{i \neq j} \left\| \mathbf{I}_i \circ \mathbf{p}_i - \mathbf{I}_j \right\|_2^2, \qquad (3)$$

where $\mathbf{p} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{N-1}\}$ is the set of transformations to apply to $\{\mathbf{I}_i\}_{i \neq j}$ to map them onto the reference $\mathbf{I}_j$. The main advantage of LSC over low-rank/entropy-based ones is faster convergence, as the adoption of the least-squares cost function allows for the use of standard Gauss-Newton optimisation techniques. On the other hand, alignment performance tend to be worse due to its simplicity.

2.1. Proposed Method

Motivated by the need for highly accurate alignment in very large-scale problems, we propose a congealing framework that leverages the advantages of adopting a similarity-based cost function (i.e. direct, such as ℓ_2-norm in LSC) *and* a complexity-based one (i.e. indirect, such as rank-based used in RASL). To perform this task *at scale*, we formulate the congealing problem in a way that can be efficiently optimised via standard back-propagation and SGD.

Our formulation can be interpreted in terms of lossy compression optimisation [44]:

$$\underset{\{\mathbf{p}_i\}_{i=1}^N}{\arg\min} \, \mathcal{D}(\mathbf{I}_{i \neq j} \circ \mathbf{p}_{i \neq j}, \mathbf{I}_j) + \lambda \, \mathcal{C}(\mathbf{I}_i \circ \mathbf{p}_i), \qquad (4)$$

where the *distortion* $\mathcal{D}$ reflects the total error when approximating the reference image $\mathbf{I}_j$ (i.e., original data) with the aligned image $\mathbf{I}_i \circ \mathbf{p}_i$ (i.e., compressed data), the *complexity* $\mathcal{C}$ is the total number of *symbols* required to encode $\mathbf{I}_i \circ \mathbf{p}_i$, and the parameter $\lambda \geq 0$ trades off the two quantities. A good candidate for our distortion (or similarity) measure $\mathcal{D}$ should be robust to occlusions, noise, outliers and, in general, objects that might partially differ in appearance (e.g. same digit but different font, same face but wearing glasses or not). In RASL, this is achieved by adding explicit constraints on the noise model and its sparsity properties which have a significant impact on the optimisation efficiency. To circumvent this problem, we adopt the ℓ_1-norm as measure of distortion, which can be efficiently optimised and offers a higher level of robustness compared to the ℓ_2-norm used in LSC or [44]. Formally,

$$\mathcal{D} = \sum_{i \neq j} \left\| \mathbf{I}_i \circ \mathbf{p}_i - \mathbf{I}_j \right\|_1. \qquad (5)$$

Motivated by the need for optimisation in large-scale joint alignment problems, we propose an efficient alternative to rank minimisation. Specifically, we observe that when a set of images are well-aligned, they form a sequence that contains a significant level of redundant information. As a

consequence, the stack of images can be compressed with higher compression rates w.r.t. the original misaligned ones. Alternatively, a lower reconstruction error can be attained, at parity of compression rate. Exploiting this consideration we therefore propose to optimise:

$$\operatorname*{argmin}_{\{\mathbf{p}_i\}_{i=1}^{N},\phi,\theta} \sum_i \left\| D_\phi(E_\theta(\mathbf{I}_i \circ \mathbf{p}_i)) - \mathbf{I}_i \circ \mathbf{p}_i \right\|_1,$$
$$\text{s.t.} \quad f(E_\theta(\mathbf{I}_i \circ \mathbf{p}_i)) \leq \beta \tag{6}$$

where ℓ_1-norm is preferred to typical ℓ_2-norm for similar reasons as the ones mentioned above; $E_\theta := \mathbb{R}^{w \times h} \to \mathbb{R}_+^b$ defines an encoder mapping an $w \times h$ image into a code vector $\mathbf{z}$ with b positive components; $D_\phi := \mathbb{R}_+^b \to \mathbb{R}^{w \times h}$ defines a decoder mapping a code $\mathbf{z}$ into a $w \times h$ image; $f := \mathbb{R}_+^b \to \mathbb{R}$ defines a (monotonically increasing) positional weighting penalty applied to $\mathbf{z}$. This penalty explicitly encourages the encoder-decoder to represent the aligned images using *primarily* the first components of $\mathbf{z}$. Similarly, β can be interpreted as an hyper-parameter controlling the *number* of first components used to represent each image, hence the representational power (or *capacity*) of the encoder-decoder block. Intuitively, at parity of encoder-decoder capacity, improving the joint alignment (i.e., optimising w.r.t. $\mathbf{p}$) will lead to increased redundancy across the image stack. In fact, we would have very similar colour intensities at the same pixel location across the image stack. Therefore, this capacity will be diverted from modelling inter-image pixel intensity distributions (merely due to misalignment) to modelling the key details of the object these images share, hence leading to lower reconstruction error. With the aim of solving large-scale alignment problems efficiently, we leverage principles from Lagrangian relaxation and penalty functions [25, 41] to approximate the solution of the constrained problem in Eq. (6) and instead propose to minimise:

$$\mathcal{C} = \sum_i \left\| D_\phi(E_\theta(\mathbf{I}_i \circ \mathbf{p}_i)) - \mathbf{I}_i \circ \mathbf{p}_i \right\|_1 + \gamma\, f(E_\theta(\mathbf{I}_i \circ \mathbf{p}_i)),$$
$$\tag{7}$$

where $\gamma \geq 0$ trades off the contribution of the reconstruction error and the capacity of the encoder-decoder block, and $\mathcal{C}$ is our measure of complexity. Plugging Eq. (5) and Eq. (7) in Eq. (4), we obtain a novel formulation to solve congealing:

$$\operatorname*{argmin}_{\{\mathbf{p}_i\}_{i=1}^{N},\phi,\theta} \sum_i \left\| \mathbf{I}_{i \neq j} \circ \mathbf{p}_{i \neq j} - \mathbf{I}_j \right\|_1 +$$
$$+ \lambda \left(\left\| D_\phi(E_\theta(\mathbf{I}_i \circ \mathbf{p}_i)) - \mathbf{I}_i \circ \mathbf{p}_i \right\|_1 + \gamma\, f(E_\theta(\mathbf{I}_i \circ \mathbf{p}_i)) \right).$$
$$\tag{8}$$

To take advantage of efficient back-propagation and SGD optimisation, (i) we implement E_θ and D_ϕ as Neural Networks (NNs) to form a low-capacity auto-encoder (controlled by γ); (ii) we define

$f(E_\theta(\mathbf{I}_j \circ \mathbf{p}_j)) \triangleq f(\mathbf{z}_j) = \mathbf{w}^\top \mathbf{z}_j$, and each component w_l of the weighing vector $\mathbf{w} = [w_1, \ldots, w_b]^\top$ is such that $w_l = l^k / \sum_{l=1}^{b} l^k$ with $k \in \mathbb{N}$; and (iii) we adopt the state-of-the-art Densely fused Spatial Transformer Network (DeSTNet) [3] as the module learning and applying the set of global transformations ($\mathbf{p}$) to the stack of images. Fig. 2 shows the proposed method for large-scale congealing. Each input image in a batch[1] is first aligned to the reference $\mathbf{I}_j$ by the DeSTNet, and the alignment error as computed by the similarity-based loss $\mathcal{D}$ is directly back-propagated to update DeSTNet's parameters to achieve better alignment to the reference. Once a batch of images has been aligned, it goes to the penalised auto-encoder: the reconstruction error as computed by $\mathcal{C}$ is used to update (i) the auto-encoder, i.e. to improve reconstruction at parity of alignment quality, *and* (ii) to further update the DeSTNet, i.e. to improve reconstruction by better alignment at parity of auto-encoder capacity. Importantly, our approach does not require gradient adjustment, as the gradient of the total loss (Eq. (8)) w.r.t. the learnable parameters is implicitly and seamlessly distributed to each module (auto-encoder and alignment), by chain-rule.

3. Experiments

We extensively evaluate the performance of the proposed method and compare it with state-of-the-art approaches [35, 15, 32] in terms of *alignment quality*, *scalability* and *robustness to noise* on MNIST [24] and several variants. To quantify performance, we adopt the *Alignment Peak Signal to Noise Ratio*, $\text{APSNR} = 10 \log_{10}\left(\frac{255^2}{\text{MSE}}\right)$ [35, 15, 32] where,

$$\text{MSE} = \frac{1}{Nhw} \sum_{i=1}^{N} \sum_{r=1}^{h} \sum_{c=1}^{w} \left(\widehat{\mathbf{I}}_i^0(r, c) - \bar{\mathbf{I}}^0(r, c) \right)^2, \tag{9}$$

$\widehat{\mathbf{I}}_i^0$ represents image i and $\bar{\mathbf{I}}^0$ the average image, both computed after alignment. We then investigate the impact of each individual term of the loss ($\mathcal{D}$ and $\mathcal{C}$) on the alignment quality and how they interact to achieve an improved level of performance when combined. With the aim of comparing the proposed method with *Deep Congealing* (DC) [16][2] and to assess the possibility of adopting the proposed method on more challenging datasets, we scale the framework and use it to jointly align multiple subsets of the LFW [18], under different initialisation.

3.1. MNIST

With the aim of evaluating the scalability of the proposed method and the baselines, we start by cre-

---

[1]We use batch-based optimisation.

[2]A comparison on MNIST and variants thereof was not possible as, to the best of our knowledge, the authors have not made the original implementation available.

Table 1. Architectures used for MNIST and LFW experiments. convD1-D2: convolution layer with D1×D1 receptive field, D2 channels, $\mathcal{F}$: fusion operation used in DeSTNet for fusing the parameters update, $|\mathbf{z}|$: dimentionality of $\mathbf{z}$. Default stride for convD1-D2 is 1, * corresponds to 2.

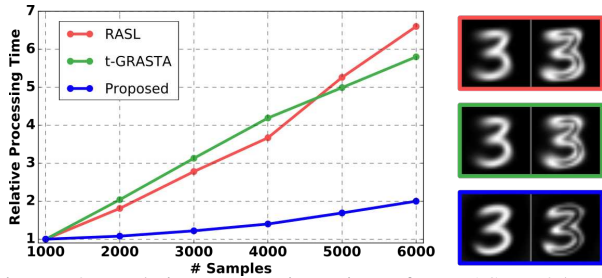|	MNIST	LFW
Aligner	$\mathcal{F}\{[$ conv7-4 $\mid$ conv7-8 $\mid$ conv1-8 $]\times 4\}$	$\mathcal{F}\{[$ conv3-64* $\mid$ conv3-128* $\mid$ conv3-256* $\mid$ conv1-8 $]\times 5\}$
Encoder	$[$conv3-100$^*] \times 3 \mid [$conv1-1024$] \times 2 \mid$ conv1-$|\mathbf{z}|$	$[$conv3-128$^*] \times 3 \mid [$conv1-512$] \times 2 \mid$ conv1-$|\mathbf{z}|$
Decoder	$[$conv1-1024$] \times 2 \mid$ conv1-16 $\mid [$conv3-100$^*] \times 3 \mid$ conv1-1	$[$conv1-512$] \times 2 \mid$ conv1-3072 $\mid [$conv3-128$^*] \times 3 \mid$ conv1-3

Figure 3. Relative processing time for RASL [35], t-GRASTA [15], and the proposed method when aligning an increasingly large number of images. Mean and variances of the aligned images produced by the compared methods for the 6 000 samples are also displayed.
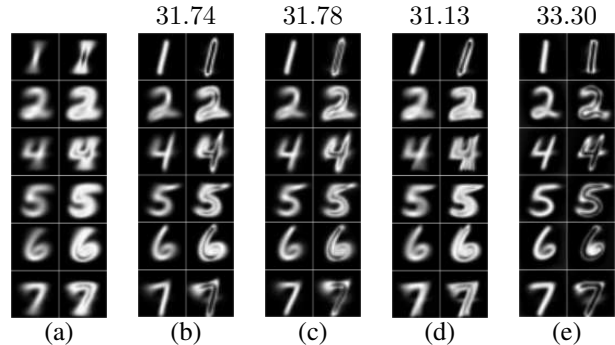

Figure 4. Congealing results on 6 000 images per digit from MNIST. (a) Before alignment, (b) RASL [35], (c) t-GRASTA [15], (d) PSSV [32], and (e) Proposed method. In each subfigure (a)-(e), the first column shows means, whereas the second one shows variances. APSNR for each digit is reported at the top of each subfigure.

ating multiple MNIST subsets, as follows. For each digit in $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, we randomly sample $\{1\,000, 2\,000, 3\,000, 4\,000, 5\,000, 6\,000\}$ images from the original MNIST dataset and align them separately. For the proposed method, we adopt *DeSTNet-4* [3] with expansion rate $k^F = 32$ as aligner, and the penalised reconstruction auto-encoder defined in Table 1, where we use *tanh* non-linearities after each layer, apart from the last layer of the encoder, where *sigmoid* is used, to keep each component of $\mathbf{z}$ in $[0, 1]$. We set $\lambda = 1$ to use both similarity- and complexity-based loss, $\gamma = 1$ and $k = 1$. We optimise the entire architecture end-to-end, using a standard Adam-based SGD optimiser with learning rate 10^{-5}. Following [35, 15, 23], we qualitatively assess alignment results for the proposed method and the baselines by computing the mean and variance across the entire dataset before and after alignment. To evaluate *scalability*, we measure the relative processing time for RASL, t-GRASTA, and the proposed method when aligning an increasingly large number of images. Due to the difference in hardware (CPUs used by the baselines, GPUs by the proposed method), we normalise processing times w.r.t. the time required to align $1, 000$ images to provide a fair comparison. As Fig. 3 shows for the case of digit '3'[3], the proposed method scales better than the baselines. Moreover, as Fig. 4 shows in the most challenging case, i.e. datasets with $6, 000$ images, the much sharper mean and lower variance images (hence higher APSNR) suggest that proposed method achieves much better alignment too. Following the experimental protocol in [27, 3],

---

[3]Similar results hold for the other digits.

we evaluate the robustness of each method to synthetic distortions based on random perspective warps. Specifically, assuming each MNIST image is $s \times s$ pixels ($s = 28$), the four corners of each image are independently and randomly scaled with Gaussian noise $\mathcal{N}(0, \sigma^2 s^2)$, then randomly translated with the same noise model. We assess alignment quality under three levels of perturbation, i.e. $\sigma = \{10\%, 20\%, 30\%\}$. To this aim, we apply this perturbation model to each $6\,000$ images dataset and report a subset of the results in Fig. 4. We observe that although a 10% perturbation seems to be well handled by RASL and t-GRASTA, alignment performance deteriorates significantly at 20% and they tend to fail at the most challenging 30%. On the other hand, the proposed method shows strong robustness to this perturbation model across all the digits and under significant noise.

3.2. Ablation Study

The proposed congealing approach takes advantage of both the similarity- and complexity-based losses (i.e., $\mathcal{D}$ and $\mathcal{C}$ in Eq. (5) and Eq. (7), respectively), as described in Eq. (8). With the aim of disentangling the contribution of each term to the final result, we have evaluated the joint alignment performance when one of the two losses is excluded from the optimisation. Figs. 6(b) and (c) show the alignment results when excluding $\mathcal{D}$, and $\mathcal{C}$, respectively, while the alignment results produced when both are used are

	31.91	32.01	31.40	33.28

$\sigma = 10\%$

	31.66	30.64	30.48	33.20

$\sigma = 20\%$

	29.80	29.36	29.47	32.96

$\sigma = 30\%$

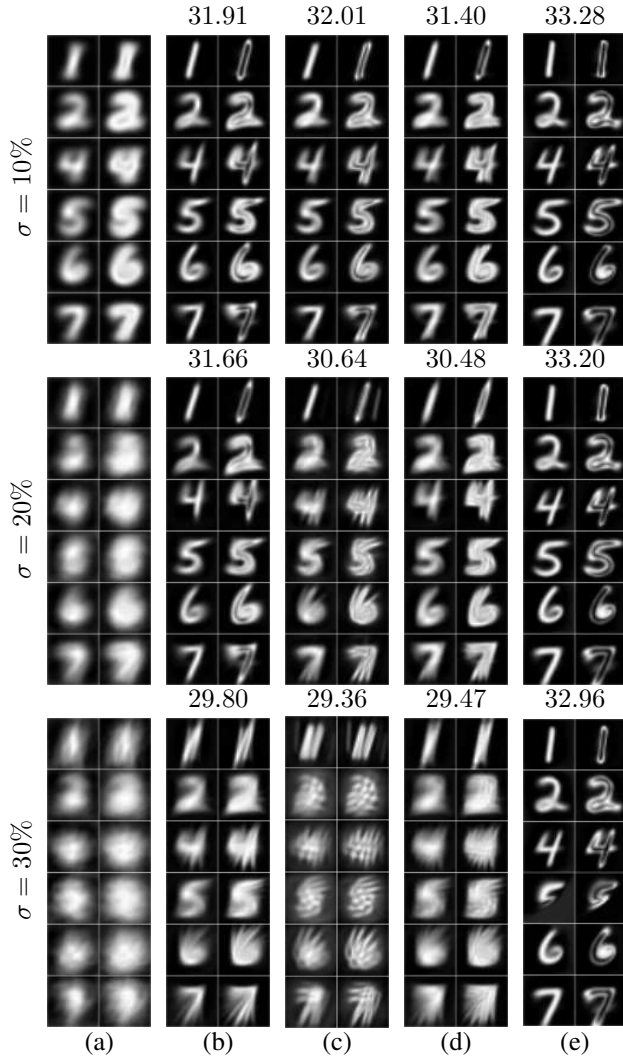(a)   (b)   (c)   (d)   (e)

Figure 5. Robustness of congealing methods to random perspective warps with $\sigma = \{10\%, 20\%, 30\%\}$, corresponding to top, middle and bottom block, respectively. (a) Before alignment, (b) RASL [35], (c) t-GRASTA [15], (d) PSSV [32], and (e) Proposed method. In each subfigure (a)-(e), the first column shows means, whereas the second one shows variances. For compactness, APSNR for each method is averaged across the digits and reported at the top of each cell.

displayed in Fig. 6(d). We observe that, in general, excluding $\mathcal{D}$ has a stronger impact on the final alignment results; moreover, the use of the reference image when computing $\mathcal{D}$ makes the optimisation much more robust, as it implicitly avoids the shrinking effect typically observed when only $\mathcal{C}$ is used. The latter is due to the fact that, at parity of reconstruction capacity for the auto-encoder, a lower complexity measure is attained when the object to reconstruct shows less spatial variability and can therefore be better re-
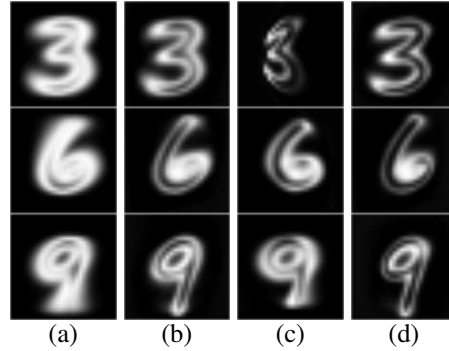

(a)   (b)   (c)   (d)

Figure 6. Ablation study: disentangling the impact of the similarity- ($\mathcal{D}$) and complexity-based ($\mathcal{C}$) losses on the final alignment result. Variance images (a) before alignment, (b) $\mathcal{D}$-only, (c) $\mathcal{C}$-only, and (d) both.

constructed[4] (see Eq. (7)). We observe that, (i) the addition of $\mathcal{C}$ to the loss based only on $\mathcal{D}$, contributes to further refining the alignment results and achieving even lower variance (see digit '6' and '9'); (ii) importantly, $\mathcal{C}$ tends to drive the overall optimisation towards solutions that favour a more (spatially) uniform alignment, as shown for digit '3'; in this sense, the complexity-based loss can be interpreted as a regulariser.

3.3. affNIST

Previously proposed congealing approaches have shown limitations in terms of scaling efficiency; in fact, on very low-resolution datasets, joint alignment optimisation results have been reported only for up to a few thousands samples [10]. Moreover, as confirmed in the experiments reported in the previous section, large intra-class spatial variability (modelled with synthetic perturbation) seems to significantly deteriorate performance. To further push the limits and evaluate the performance of the proposed method, we assess joint alignment performance on a much more challenging version of MNIST, namely *affNIST* [43]. This dataset is built by taking images from MNIST and applying various reasonable affine transformations to them. In the process, the images become 40×40 pixels large, with significant translations involved. From this dataset, we take the first $100\,000$ samples for each digit and perform alignment (results in Fig. 7), using the same parameter setting adopted in the experiments above. The strong variability characterising this dataset is clear by looking at the means and variances before alignment, and a subset of the actual inputs (Fig. 7-middle). Nevertheless, the proposed method achieves a good level of alignment, as demonstrated by the average and variance images after alignment (hence high APSNR) and a subset of the actual outputs (Fig. 7-bottom).

---

[4]Notice that this undesired effect is typical of low-rank-based congealing approaches [35].

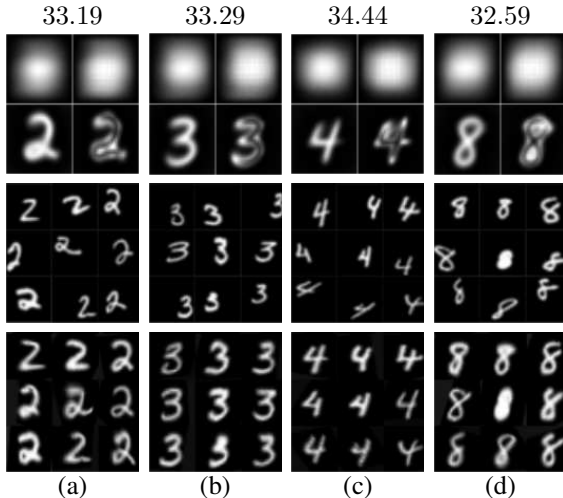| 33.19 | 33.29 | 34.44 | 32.59 |

| (a) | (b) | (c) | (d) |

Figure 7. Congealing results of the proposed method on $100\,000$ images per digit from affNIST. (a)-(d) correspond to different digits. Top: mean (first columns) and variance (second columns) images, before (first rows) and after (second rows) alignment. Middle: a subset of the actual inputs. Bottom: a subset of the actual outputs. APSNR for each digit is reported at the top of each subfigure.

| 32.20 | 32.63 | 31.93 |

| (a) | (b) | (c) |

Figure 8. Congealing results of the proposed method on $1\,000\,000$ images per digit from infiMNIST. (a)-(c) correspond to different digits. Top: mean (first columns) and variance (second columns) images, before (first rows) and after (second rows) alignment. Middle: a subset of the actual inputs. Bottom: a subset of the actual outputs. APSNR for each digit is reported at the top of each subfigure.

3.4. infiMNIST

So far, the proposed method has shown robustness to global affine/perspective perturbations, and joint alignment problems with up to $100,000$ samples per digit. Here, we evaluate the alignment performance under non-linear (local) deformations (e.g. tickening) and translations, and solve the joint alignment problem for $1,000,000$ images per digit sampled from infiMNIST [28][5]. Notice, we use the same parameter setting adopted above to assess the robustness and generalisation of the proposed method in a much more challenging joint alignment problem. As Fig. 8 shows, despite the random translations being relatively smaller than the ones used in affNIST, the non-linear perturbations add a much higher level of intra-class variability. Nevertheless, the proposed method achieves remarkable joint alignment at this scale and under this kind of perturbations.

3.5. LFW

LFW [22] has been widely used to assess the performance of state-of-the-art joint alignment methods, e.g. in [35, 16]. This dataset is made challenging by multiple factors, including variations in facial expression, occlusion, illumination changes, clutter in the background and head pose variations. Moreover, each subject image is 250×250 pixels, which is much larger than MNIST (and variants) images used in the experiments above. We selected four subsets, corresponding to male and fe-

male subjects with the largest amount of images, namely `George_W_Bush`, `Tony_Blair`, `Serena_Williams`, and `Jennifer_Capriati`. To accommodate the difference in input image size and considering the more complex task w.r.t. MNIST-based datasets, we scale the aligner and the encoder-decoder block as reported in Table 1. In Fig. 9, we report a qualitative and quantitative comparison of the proposed method with RASL [35], PSSV [32] and Deep Congealing [16], for which joint alignment results initialised with the Viola-Jones face detector [45] are available at http://vis-www.cs.umass.edu/lfw/. For fair comparison, we adopt the same initialisation for the proposed method and the baselines. We observe that, overall, the proposed method outperforms both RASL, PSSV and Deep Congealing, in terms of APSNR which is qualitatively confirmed by sharper average images across all the subjects. Moreover, unlike RASL and PSSV, the proposed method does not suffer a zoom-in/zoom-out effect which makes the optimisation focus on smaller/larger portion of the region of interest. This can be attributed to the use of the reference image in $\mathcal{D}$.

Although important progress has been made in recent years in face detection [7, 46, 48, 42], some level of inaccuracy is inevitable in a practical setting. So, it is important to assess the robustness of the proposed method to coarser initialisation. To this aim, we increased the size of the initial bounding box returned by the Viola-Jones face detector by 15% and 30% in width and height, and report the joint alignment results in Fig. 10. We observe that the performance of both RASL (Figs 10(b,e)) and PSSV (Figs 10(c,f)) degrade

---

[5]The code to generate datasets from infiMNIST is available at https://leon.bottou.org/projects/infiMNIST.

29.90   29.74   30.31   30.73
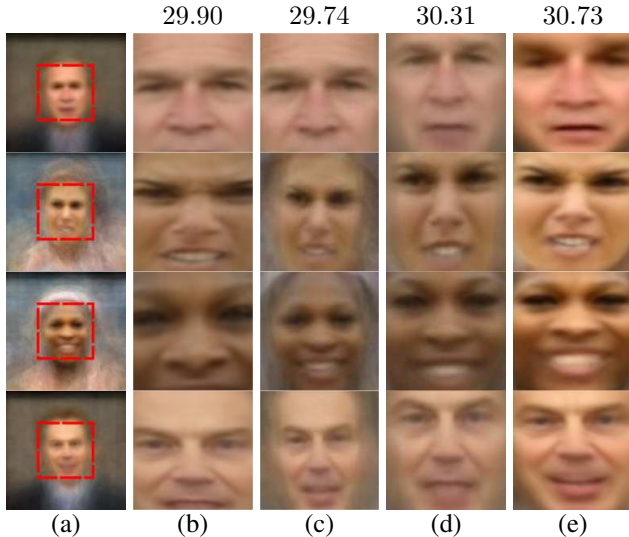
(a)   (b)   (c)   (d)   (e)

Figure 9. Congealing results (means) on LFW. (a) Before alignment, (b) RASL [35], (c) PSSV [32], (d) Deep Congealing [16], and (e) Proposed method. The bounding box initialisation is shown in red in (a) for all the subjects. For compactness, average APSNR for each method is reported at the top of each subfigure and averaged across the subjects.

Init [45]   ←— +15% —→   ←— +30% —→

29.75   29.25   30.43   29.14   28.67   30.29
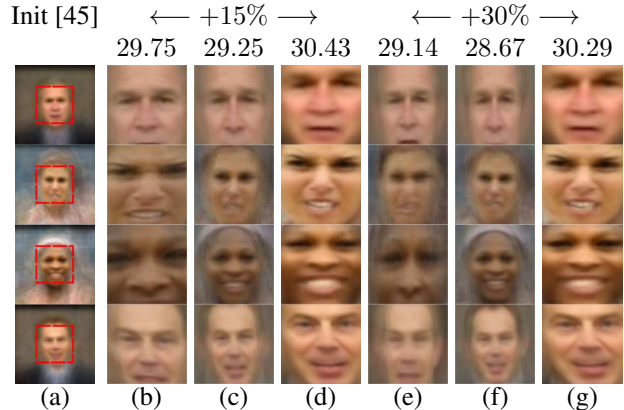
(a)   (b)   (c)   (d)   (e)   (f)   (g)

Figure 10. Robustness of congealing methods to initialisation, i.e. bounding box 15% and 30% larger than the one estimated by [45] (in red colour in (a)). Mean images (a) before alignment, (b)(e) RASL [35], (c)(f) PSSV [32], and (d)(g) Proposed method. For compactness, average APSNR for each method is reported at the top of each subfigure and averaged across the subjects.

significantly when the initialisation is not close to the object, as confirmed by a sharper decrease in average APSNR and the average aligned faces being blurry. Instead, the proposed method demonstrates strong robustness to the initialisation: as can be observed in Figs 10(d,g), our mean aligned faces are clean and crisp which indicates a remarkable level of alignment even with a bounding box 30% larger.

Following the protocol adopted in [35, 15], we further quantify alignment performance by computing the average errors in the locations of three landmarks (the eye outer corners and tip of nose), calculated as the distances of the estimated locations to their centre, normalised by the eye-to-eye distance. We compare our alignment performance against RASL (best rank-based baseline) and DC (deep learning approach). We average the performance for each landmark in a given subject and report them in Table 2. Confirming the considerations above, when the original initialisation is adopted, the proposed method attains the lowest errors across all the subjects. Moreover, while at 15% coarser initialisation RASL starts to show difficulties on some subjects, at 30% performance degrades significantly. Instead, the proposed method shows much stronger robustness across subjects and initialisation.

4. Conclusions

Image alignment is a major area of research in computer vision. However, the majority of previously proposed methods focus on identifying pixel-level correspondences between a *pair* of images. Instead, a plethora of

Table 2. Average errors for three landmarks (the eye outer corners and tip of nose), calculated as the distances of the estimated locations to their centre, normalised by the eye-to-eye distance. S1:George_W_Bush, S2:Jennifer_Capriati, S3:Serena_Williams, S4:Tony_Blair.

Init	Methods	S1	S2	S3	S4
[45]	RASL [35]	2.88%	2.45%	3.32%	3.24%
	DC [16]	3.97%	3.48%	3.48%	3.27%
	Proposed	**2.67%**	**1.86%**	**2.24%**	**2.39%**
+15%	RASL [35]	**3.24%**	6.40%	5.02%	3.65%
	Proposed	3.84%	**2.12%**	**4.34%**	**2.04%**
+30%	RASL [35]	6.29%	6.77%	7.08%	6.87%
	Proposed	**4.27%**	**1.92%**	**3.69%**	**2.55%**

other tasks such as, co-segmentation, image edit propagation and structure-from-motion, would considerably benefit from establishing pixel-level correspondences between a *set* of images. Several congealing or joint alignment methods have been previously proposed; however, scalability to large datasets and the limited robustness to initialisation and intra-class variability seem to hamper their wide applicability. To address these limitations, we have proposed a novel congealing method and shown that it is capable of handling joint alignment problems at very large scale i.e., up to one million data points, simultaneously. This is achieved through a novel *differentiable* formulation of the congealing problem, which combines the advantages of similarity- and rank-based congealing approaches and can be easily optimised with standard SGD, end-to-end. Extensive experimental results on several benchmark datasets, including digits and faces at different resolutions, show that the proposed congealing framework outperforms state-of-the-art approaches in terms of scalability, alignment quality and robustness to linear and non-linear geometric perturbations of different magnitude and type.

References

[1] Roberto Annunziata, Ahmad Kheirkhah, Shruti Aggarwal, Pedram Hamrah, and Emanuele Trucco. A fully automated tortuosity quantification system with application to corneal nerve fibres in confocal microscopy images. *Medical image analysis*, 32:216–232, 2016.

[2] Roberto Annunziata, Ahmad Kheirkhah, Pedram Hamrah, and Emanuele Trucco. Scale and curvature invariant ridge detector for tortuous and fragmented structures. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 588–595. Springer, 2015.

[3] Roberto Annunziata, Christos Sagonas, and Jacques Calì. DeSTNet: Densely fused spatial transformer networks. *Proceedings of British Machine Vision Conference (BMVC)*, 2018.

[4] Roberto Annunziata and Emanuele Trucco. Accelerating convolutional sparse coding for curvilinear structures segmentation by refining SCIRD-TS filter banks. *IEEE Transactions on Medical Imaging (TMI)*, 35(11):2381–2392, 2016.

[5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 404–417, 2006.

[6] Ryoma Bise, Yingqiang Zheng, Imari Sato, and Masakazu Toi. Vascular registration in photoacoustic imaging by low-rank alignment via foreground, background and complement decomposition. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 326–334, 2016.

[7] Dong Chen, Gang Hua, Fang Wen, and Jian Sun. Supervised transformer network for efficient face detection. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 122–138, 2016.

[8] Xiai Chen, Zhi Han, Yao Wang, Yandong Tang, and Haibin Yu. Nonconvex plus quadratic penalized low-rank and sparse decomposition for noisy image alignment. *Science China Information Sciences*, 59(5):052107, 2016.

[9] Mark Cox, Sridha Sridharan, Simon Lucey, and Jeffrey Cohn. Least squares congealing for unsupervised alignment of images. In *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, page 1, 2008.

[10] Mark Cox, Sridha Sridharan, Simon Lucey, and Jeffrey Cohn. Least-squares congealing for large numbers of images. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 1949–1956, 2009.

[11] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, volume 1, pages 886–893, 2005.

[12] David A Forsyth and Jean Ponce. A modern approach. *Computer vision: a modern approach*, pages 88–101, 2003.

[13] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[14] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[15] Jun He, Dejiao Zhang, Laura Balzano, and Tao Tao. Iterative grassmannian optimization for robust image alignment. *Image and Vision Computing (IMAVIS)*, 32(10):800–813, 2014.

[16] Gary Huang, Marwan Mattar, Honglak Lee, and Erik G Learned-Miller. Learning to align from scratch. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[17] Gary B Huang, Vidit Jain, and Erik Learned-Miller. Unsupervised joint alignment of complex images. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

[18] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in Real-Life Images: detection, alignment, and recognition*, 2008.

[19] Juan Eugenio Iglesias, Marc Modat, Loïc Peter, Allison Stevens, Roberto Annunziata, Tom Vercauteren, Ed Lein, Bruce Fischl, Sebastien Ourselin, Alzheimers Disease Neuroimaging Initiative, et al. Joint registration and synthesis using a probabilistic model for alignment of mri and histological sections. *Medical image analysis*, 50:127–144, 2018.

[20] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568. ACM, 2011.

[21] Angjoo Kanazawa, David W Jacobs, and Manmohan Chandraker. Warpnet: Weakly supervised matching for single-view reconstruction. In *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 3253–3261, 2016.

[22] Erik Learned-Miller, Gary B Huang, Aruni RoyChowdhury, Haoxiang Li, and Gang Hua. Labeled faces in the wild: A survey. In *Advances in face detection and facial image analysis*, pages 189–248. Springer, 2016.

[23] Erik G Learned-Miller. Data driven image models through continuous joint alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(2):236–250, 2006.

[24] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[25] Claude Lemaréchal. Lagrangian relaxation. In *Computational combinatorial optimization*, pages 112–156. Springer, 2001.

[26] Habte Tadesse Likassa, Wen-Hsien Fang, and Yung-An Chuang. Modified robust image alignment by sparse and low rank decomposition for highly linearly correlated data. In *2018 3rd International Conference on Intelligent Green Building and Smart Grid (IGBSG)*, pages 1–4. IEEE, 2018.

[27] Chen-Hsuan Lin and Simon Lucey. Inverse compositional spatial transformer networks. In *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 2568–2576, 2017.

[28] Gaëlle Loosli, Stéphane Canu, and Léon Bottou. Training invariant support vector machines using selective sampling. *Large scale kernel machines*, 2, 2007.

[29] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.

[30] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJAI)*, pages 1568–1575, 1981.

[31] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Largescale image retrieval with attentive deep local features. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 3456–3465, 2017.

[32] Tae-Hyun Oh, Yu-Wing Tai, Jean-Charles Bazin, Hyeong-woo Kim, and In So Kweon. Partial sum minimization of singular values in robust pca: Algorithm and applications. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):744–758, 2015.

[33] Zhaoqing Pan, Yun Zhang, and Sam Kwong. Efficient motion and disparity estimation optimization for low complexity multiview video coding. *IEEE Transactions on Broadcasting*, 61(2):166–176, 2015.

[34] Xi Peng, Shaoting Zhang, Yu Yang, and Dimitris N Metaxas. Piefa: Personalized incremental and ensemble face alignment. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 3880–3888, 2015.

[35] Yigang Peng, Arvind Ganesh, John Wright, Wenli Xu, and Yi Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(11):2233–2246, 2012.

[36] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[37] Christos Sagonas, Yannis Panagakis, Stefanos Zafeiriou, and Maja Pantic. Raps: Robust and efficient automatic construction of person-specific deformable models. In *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 1789–1796, 2014.

[38] Christos Sagonas, Yannis Panagakis, Stefanos Zafeiriou, and Maja Pantic. Robust statistical face frontalization. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 3871–3879, 2015.

[39] Christos Sagonas, Yannis Panagakis, Stefanos Zafeiriou, and Maja Pantic. Robust statistical frontalization of human and animal faces. *International Journal of Computer Vision (IJCV)*, 122(2):270–291, 2017.

[40] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 118–126, 2015.

[41] Alice E Smith and David W Coit. Penalty functions. *Handbook of Evolutionary Computation*, 97(1):C5, 1995.

[42] Xudong Sun, Pengcheng Wu, and Steven CH Hoi. Face detection using deep learning: An improved faster rcnn approach. *Neurocomputing*, 299:42–50, 2018.

[43] Tijmen Tieleman. Affnist. *https://www.cs.toronto.edu/ tijmen/affNIST/*.

[44] Andrea Vedaldi, Gregorio Guidi, and Stefano Soatto. Joint data alignment up to (lossy) transformations. In *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, 2008.

[45] Paul Viola and Michael J Jones. Robust real-time face detection. *International Journal of Computer Vision (IJCV)*, 57(2):137–154, 2004.

[46] Stefanos Zafeiriou, Cha Zhang, and Zhengyou Zhang. A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding (CVIU*, 138:1–24, 2015.

[47] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 4353–4361, 2015.

[48] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.