Adversarial training with cycle consistency for unsupervised super-resolution in endomicroscopy

Daniele Ravì [a,*], Agnieszka Barbara Szczotka [b], Stephen P Pereira [c], Tom Vercauteren [d]

[a] Centre for Medical Image Computing, University College London, United Kingdom
[b] Wellcome/EPSRC Centre for Interventional and Surgical Sciences, University College London, United Kingdom
[c] Institute for Liver and Digestive Health, University College London, United Kingdom
[d] School of Biomedical Engineering & Imaging Sciences, King's College London, United Kingdom

ARTICLE INFO

ABSTRACT

In recent years, endomicroscopy has become increasingly used for diagnostic purposes and interventional guidance. It can provide intraoperative aids for real-time tissue characterization and can help to perform visual investigations aimed for example to discover epithelial cancers. Due to physical constraints on the acquisition process, endomicroscopy images, still today have a low number of informative pixels which hampers their quality. Post-processing techniques, such as Super-Resolution (SR), are a potential solution to increase the quality of these images. SR techniques are often supervised, requiring aligned pairs of low-resolution (LR) and high-resolution (HR) images patches to train a model. However, in our domain, the lack of HR images hinders the collection of such pairs and makes supervised training unsuitable. For this reason, we propose an unsupervised SR framework based on an adversarial deep neural network with a physically-inspired cycle consistency, designed to impose some acquisition properties on the super-resolved images. Our framework can exploit HR images, regardless of the domain where they are coming from, to transfer the quality of the HR images to the initial LR images. This property can be particularly useful in all situations where pairs of LR/HR are not available during the training. Our quantitative analysis, validated using a database of 238 endomicroscopy video sequences from 143 patients, shows the ability of the pipeline to produce convincing super-resolved images. A Mean Opinion Score (MOS) study also confirms this quantitative image quality assessment.

1. Introduction

According to a recent report by the World Health Organization, cancer is the second leading cause of death after cardiovascular disease and was responsible for 8.8 million deaths in 2015. Early detection, such as the ability to detect precancerous lesions, plays an important role in reducing cancer incidence and related mortality (Torre et al., 2016). Optical endomicroscopy, based for example on confocal microscopy, optical coherence tomography or spectroscopy, has the ability to perform optical biopsies and identify early pathology in tissues or organs including the colon, oesophagus, pancreas, brain, liver and cervix (Ravì et al., 2017; Nguyen et al., 2015). Although in the last years, progress has been made to build reliable optical endomicroscopy devices (Neumann et al., 2010), the need to operate at micron scale through the use of endoscopes, fibre bundles, laparoscopes, and needles, limits the final

resolution of the images. Further hardware improvements are difficult to achieve and one possibility to improve the image quality is to post-process the images using SR techniques.

Recent methods for SR are based on training example-based models that learn how to improve image resolution by exploiting a database of aligned pairs of LR and HR images (Ledig et al., 2017; Ravì et al., 2018). Nonetheless, due to the lack of HR endomicroscopy images, these pairs are not typically available in this domain. An option is to generate these pairs synthetically, but achieving this in a sufficiently realistic manner is only feasible when the acquisition process is extremely well defined. In most of the cases, the acquisition process is only known approximately and supervised methods may thus not be applicable.

For this reason, we designed a deep learning architecture trained in an unsupervised manner where the aforementioned one-to-one alignment between LR and HR is not required anymore. We formalize our framework so that LR images from an initial input domain I^{LR} could be transformed into images of any target domain T^{HR}. The target domain can be the same or different from

* Corresponding author.
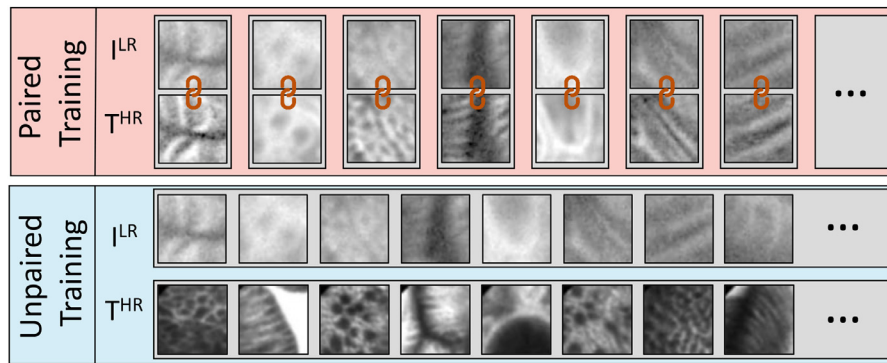*E-mail address:* d.ravi@ucl.ac.uk (D. Ravì).

Fig. 1. Example of aligned-paired and unpaired patches used for training super-resolution neural networks.

the initial one. An example of the difference between initial and target domain with paired and unpaired patches is shown in Fig. 1. To train the model using unpaired patches and avoid that the network learns to produce HR images with no direct relationship with the input images, a cycle consistency block is included in our architecture. This block imposes some physical acquisition properties so that the obtained HR images faithfully represent the initial LR images.

More specifically, during the training procedure, we make use of an adversarial network, a class of artificial intelligence algorithms which train two separate models that challenge each other in a zero-sum game. The first model is a *SR* network that learns how to improve the resolution of the images, and the second is a discriminative network *DS* that, looking at the target domain, tries to distinguish images generated by the *SR* network from the real T^{HR} images. The aim of *SR* is to learn how to fool the *DS* network and this leads to a generation of super-resolved images.

Adversarial training can learn how to produce outputs with the same distribution as the target domain. However, the target domain distribution could be sampled by simply mapping the input images to any random permutation of images in the target domain. Therefore, in this context and without specific constraints, an adversarial loss, alone, cannot guarantee that the learned function maps an input to a desired corresponding super-resolved image. Thus, following the idea proposed by Zhu et al. (2017) in the context of style transfer, we add in the adversarial training a further cycle block that imposes a consistency between the HR images and the initial LR images. In contrast to the work in Zhu et al. (2017) where the reverse mapping between the target and the source domain is also learned, in this work, this consistency is obtained by constraining the super-resolved T^{HR} image to have similar physical acquisition properties to the initial I^{LR} image.

As explained in more detail in Section 3.3, the raw signal of pCLE images is acquired from tens of thousands of fibres irregularly placed in a bundle. Moreover, the LR images are reconstructed using a Delaunay-based linear interpolation that interpolates pixels from the centres of the fibres to a regular grid. Starting from the super-resolved pCLE images created by our framework, the proposed physical constraints impose that the values obtained by inverting the aforementioned interpolation are similar to the raw signal acquired from the fibres for the corresponding LR image.

To the best of our knowledge, this paper is the first to propose an adversarial network that takes advantage of the knowledge of the physical acquisition process by imposing a cycle consistency to perform unsupervised SR of medical images. In our experiment, we show that the proposed framework does not require paired aligned patches for the training. This is an important property in all the domains where HR images are not available. The rest of the paper is organized as follows: Section 2 presents the state-of-the-art for unsupervised SR methods. Section 3 presents the proposed training

methodology based on an adversarial training with cycle consistency. Section 4 presents the results obtained using a quantitative image quality assessment and a Mean Opinion Score (MOS) study and Section 5 summarizes the contribution of this research.

2. Related work

With the recent outbreak of deep learning, example-based super-resolution (EBSR) has led to a dramatic leap in SR performance. These approaches are mainly based on a supervised training procedure where a database of aligned pairs of LR and HR images is required to create the model. Being supervised, these SR methods are restricted to specific training data, where the LR images are usually predetermined from their HR counterparts. However, in many contexts, such as in endomicroscopy, HR images are not available due to physical constraints and therefore these paired aligned images cannot be generated. A first attempt to train an EBSR network for endomicroscopy was proposed by Ravì et al. (2018) where a video-registration technique is used to estimate the HR images from a sequence of LR images. A pipeline for generation of synthetic data is finally presented to produce the desired aligned pairs. Although models trained with generated synthetic data can obtain convincing SR images, the domain gap between synthetic LR images and original pCLE images raises questions about their reliability for clinical use. For this reason, we believe that unsupervised super-resolution techniques would be more suitable in these cases. In (Ayasso et al., 2012) is presented an unsupervised method for image SR based on a Variational Bayesian (VB) algorithm that combines a Bayesian technique with a Markovian model. The main issue with this approach is the difficulty to hand-craft a good perceptual loss function and the final images tend to be blurred. Rather than designing a suitable similarity loss function, Goodfellow et al. (2014) proposed a general framework called Generative Adversarial Network (GAN) where the perceptual loss function is trained directly using a discriminative network. This allows the method to automatically verify if a generated sample is similar to a real one from the target domain. In particular, the adversarial process uses two models: i) a generative model *G*, and ii) a discriminative model *D* that are trained to play a zero-sum game. Following this general framework, Ledig et al. (2017) proposed a single image super-resolution architecture called SRGAN. Although this approach is unsupervised, part of its loss is still supervised. In fact, a content loss term based on a per-pixel loss between the output and ground-truth images is used there. This term requires again alignment between LR and HR thereby limiting its applicability in our context. Another drawback of SRGAN is its difficulty to train, often generating SR images that are too sharp or have artefacts. To reduce these drawbacks, Bao et al. (2017) proposed to combine a VB approach with GAN. They show that an asymmetric loss function obtained using

a cross-entropy loss for the discriminative network and a mean discrepancy objective for the generative network, make the GAN training more stable. Similarly to this idea, an Adversarial Variational Bayes was proposed by Mescheder et al. (2017) where a Variational Autoencoder (VAE) is trained using an auxiliary discriminative network. Contrary to the previous case, this approach provides a more clear theoretical justification. However, the problem of using paired LR/HR has not been resolved by any of the approaches described so far. One of the first approaches that formalize the possibility to translate images from a source domain X to a target domain Y in the absence of paired examples was proposed by Zhu et al. (2017) and is called CycleGAN. Using an adversarial training the goal of this method is to learn a mapping $G: X \rightarrow Y$ such that the distribution of images from $G(X)$ is indistinguishable from the distribution Y. Since this mapping is highly under-constrained, the authors also introduced an inverse mapping $F: Y \rightarrow X$ and a cycle consistency loss to ensure $F(G(X)) \approx X$. Thanks to this two-step consistency, the need for paired images is eliminated. Varying the input-output domain, this framework can be used to perform artistic style transformation (Johnson et al., 2016) (where, for example, horses can be converted into zebras) or, as in our case, transfer the resolution from one domain to another.

Some other interesting approaches were proposed by Shocher et al. (2017) and Bulat et al. (2018). Here the authors question that the predetermined LR images obtained from standard bi-cubic down-sampling rarely look like the real LR images. In Shocher et al. (2017) the authors introduce a method called Zero-Shot SR, that does not rely on prior training. To do so they exploit the internal recurrence of information inside a single image and train a small image-specific CNN at test time. This facilitates self-training SR for biological data, old photos, noisy images, and other images where the acquisition process is unknown.

Following the CycleGAN concept of Zhu et al. (2017), we propose an unsupervised framework that uses unpaired images and is designed to overcome the limitations that standard SR approaches have when aligned pairs of LR/HR images are required.

3. Materials and methods

3.1. Database

To validate our solution, we used the database proposed by André et al. (2011) containing 238 anonymized probe-based Confocal Laser Endomicroscopy (pCLE) video sequences from 143 patients captured on the colon and oesophagus regions. This database does not provide the real ground truth of the HR images and only estimated $\widehat{HR}$, computed using a time-consuming video-registration technique on the LR images are available. Video-registration may generate $\widehat{HR}$ that are not perfectly aligned with the LR and might display further reconstruction artefacts. We define this set of data as DB_{orig}. A second version of this database called DB_{syn} and based on the simulated framework proposed in Ravì et al. (2018) is also used in our experiments. In this case, the LR images are synthetically generated from the $\widehat{HR}$ and this results in paired images perfectly aligned. More specifically, the simulated framework extract fibre signals fs directly from the $\widehat{HR}$ image, by averaging the $\widehat{HR}$ pixel values within the region defined by the Voronoi cell computed from the centre of the fibre's position. Moreover, to replicate realistic noise patterns on the simulated LR images, additive and multiplicative Gaussian noise is added to each fibre signal fs obtaining a noisy fibre signal nfs. Finally, Delaunay-based linear interpolation is performed thereby leading to the final simulated LR images.

Fig. 3 shows some example of images extracted from both the two versions of the database. In both these scenarios, the database was divided randomly into three subsets: a train set (70%), a validation set (15%), and a test set (15%). The number of images that belong to each clinical setting is maintained equal in each of these set.

We provide the results from two different case studies: CS_1 where the images that belong to the same video are only distributed within one of the subsets (train, test or validation) and CS_2 where, additionally, the images from the same patient are only distributed within one of these subsets. CS_1 allows us to understand if the system is capable of super-resolve new visual patterns that have never been seen before. Given the size of our dataset, CS_2 allows for coarser but less prone-to-bias evaluation that mimics a more realistic scenario where the effectiveness of the system to transfer the inter-patient super-resolution capability to new patients is analysed. In our experimental section, we show that these two scenarios demonstrate similar trends.

Following the same pre-processed steps proposed in Ravì et al. (2018), the intensity values are first normalized to have mean zero and variance one, then the pixels values were scaled of every frame individually in the range [0–1] and finally, non-overlapping patches of 64×64 pixels were extracted only from the pCLE field of view of the train and validation set. The patches in the validation set were used to monitor the loss and avoid overfitting. Test images with size 512×512 were processed at full-size to compute the final results. We highlight that the proposed SR framework does not have the aim to increase the number of pixels, but rather to improve the quality of the LR images that are initially oversampled with an average of 7 interpolated pixels for each informative fibre-pixel. The choice to initially oversample is made by the manufacturer to ensure that the image space is discretised in a sufficiently fine manner to map the fibre graph onto a square pixel grid without too much distortion. In conclusion, in our system, the output images have the same size as the input images but display refined content. The full-size processing of the test images is possible since the inference network is fully convolutional and no specific image size is required as input.

3.2. Adversarial training

The pipeline used for training our framework is presented in Fig. 2 and is divided into different sub-sections, each coded by a specific colour.

We formalize our training as an adversarial min-max problem where two networks, a discriminative network defined as DS_γ (red sub-sequence in Fig. 2), and a super-resolution network defined as SR_θ (grey sub-sequence in Fig. 2) are trained concurrently. More
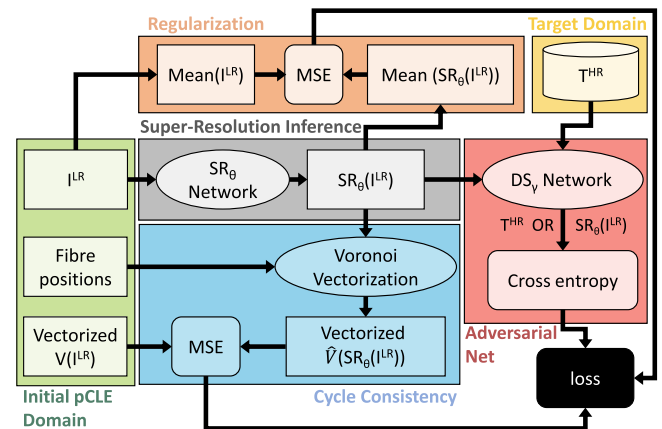

Fig. 2. Pipeline used for training the proposed adversarial network with cycle consistency. Each component of the pipeline is identified by a different colour. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
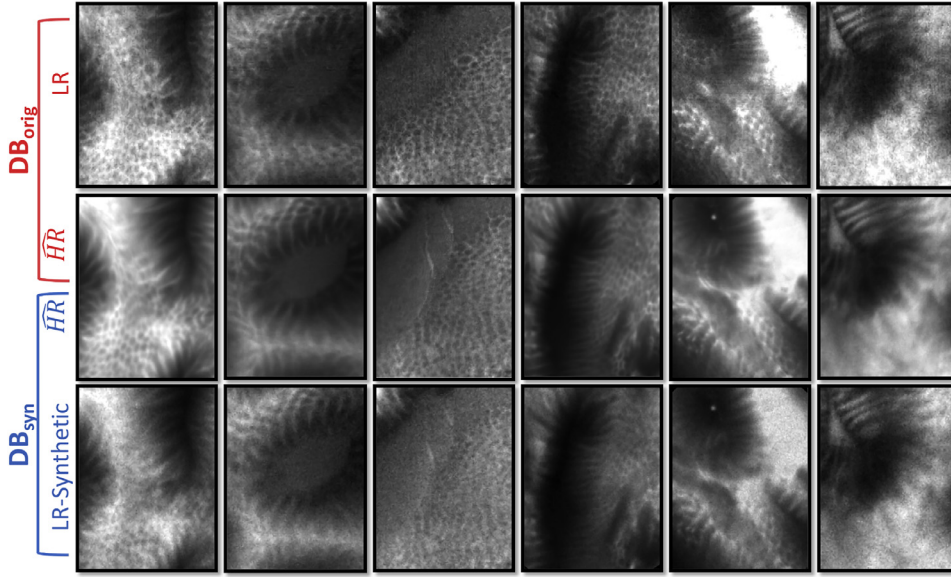
Fig. 3. Example of images contained in the two proposed versions of the database. Both the two versions share the same $\widehat{HR}$ images. The LR in DB_{orig} are images acquired on human tissues and they are not always aligned with the related $\widehat{HR}$. The LR in DB_{syn} are instead generated synthetically and are always aligned with the corresponding $\widehat{HR}$.

specifically, the first network DS_γ is trained solving:

$$\max_\gamma \mathbb{E}_{x \sim p_{ILR}} \left[log\left(1 - DS_\gamma\left(SR_\theta(x)\right)\right)\right] + \mathbb{E}_{y \sim p_{THR}} \left[log DS_\gamma(y)\right], \quad (1)$$

where p_{ILR} and p_{THR} are respectively the patch distributions on the input and target domain, $DS_\gamma(*)$ estimates the probability that a patch comes from the target domain, whereas $SR_\theta(x)$ is the predicted super-resolved patch obtained from x. The meaning of Eq. (1) is that the discriminator has to maximize how to discriminate predicted super-resolved images from real T^{HR} patches.

The second network SR_θ, is trained instead through the minimization of a composite loss function $loss_t$ obtained solving:

$$\min_\theta \mathbb{E}_{x \sim p_{ILR}} \left[loss_t\left(x, SR_\theta(x)\right)\right] \quad (2)$$

The proposed $loss_t$, defined in Eq. (3), is a combination of three terms: l_{Vec} that models the physical acquisition characteristics of the predicted super-resolved patch, l_{Adv} that models the adversarial loss function and l_{Reg} used to regularize the network training. The details of each term are provided later in this section.

$$loss_t = l_{Vec} + l_{Adv} + l_{Reg} \quad (3)$$

Both SR_θ and DS_γ are concurrently trained using the back-propagation algorithm that gradually adjusts the parameters θ and γ through a stochastic gradient descent for the former and a stochastic gradient ascent for the latter.

3.3. Input domain and cycle consistency

3.3.1. Input domain

The green blocks in Fig. 2 represent the data structures required as input for the proposed pipeline. The most obvious input is the reconstructed I^{LR} that is used by the SR_θ network to infer the super-resolved patch $SR_\theta(I^{LR})$.

In the pCLE imaging, image acquisition is achieved by illuminating one fibre at a time. Each fibre acts as an individual pinhole and a scan point for fibre confocality. The information from all the fibres is then collected in a vector that we refer to as a vectorized image $V(I^{LR})$ and represents the main input block in our pipeline. I^{LR} images are reconstructed interpolating the values in $V(I^{LR})$ from the centres of the fibre positions to the points of a regular grid.

Therefore the fibre positions are the other key input block required by our pipeline.

3.3.2. Cycle consistency

Starting from a generated high-resolution pCLE image $SR_\theta(I^{LR})$, we can obtain a low-resolution representation of it, by a process referred to as Voronoi vectorization $\widehat{V}(SR_\theta(I^{LR}))$ which is equivalent to the down-sampling for standard images. The details of the Voronoi vectorization used in our framework are described in Fig. 4. Here, the first step is to compute the Voronoi diagram from the fibre positions. The result is a partition of the plane where for each fibre there is a corresponding region, called Voronoi cell, consisting of all points closer to this fibre than to any other fibre. The next step is to average the pixels in the $SR_\theta(I^{LR})$ patch that belongs to the same Voronoi cell, imitating the point spread function of the fibre acquisition process. All the elements in the vector are then normalized in the range [0, 1]. This normalization makes the training faster and reduces the chances of getting stuck in local optima. Since each patch may have a different number of fibres, the vectorization can produce vectors of different sizes. Therefore as a final step, a 0-padding is introduced so that each vector always has a fixed number of elements. We define this fixed number as N_F that is equal to the maximum number of fibres in a single patch. In our database, N_F is 682 which is commensurate with the ratio between the patch size (64×64) and the average factor (7) used to oversample each informative fibre-pixel.

The vectorized $V(I^{LR})$ and the Voronoi vectorization $\widehat{V}(SR_\theta(I^{LR}))$ are used in our pipeline to create the cycle consistency (blocks coloured in cyan in Fig. 2). These blocks are used to impose the requirements for the predicted super-resolved images $SR_\theta(I^{LR})$ to have the same physical acquisition properties as the initial I^{LR} images. Without this cycle consistency, the network could simply produce arbitrary images in the target domain with no relationship to the structures contained in the input image, because our framework relies on unpaired patches. To avoid this, we force the $V(I^{LR})$ and $\widehat{V}(SR_\theta(I^{LR}))$ to be similar using the l_{Vec} term in the proposed loss function.

$$l_{Vec} = \frac{1}{N_F} \sum_{i=1}^{N_F} \left[V(I^{LR})_i - \widehat{V}(SR_\theta(I^{LR}))_i \right]^2 \quad (4)$$
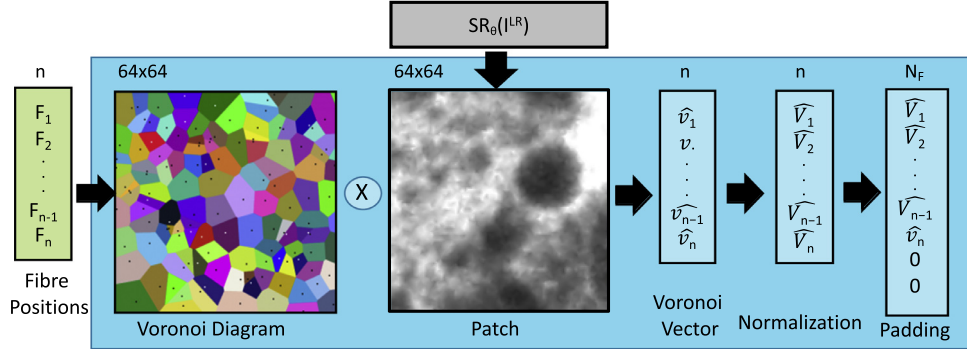
Fig. 4. Voronoi vectorization used in our pipeline to constrain the predicted super-resolved patches.

In contrast to CycleGan (Zhu et al., 2017), our cycle consistency block is not a trainable network, but rather is used to constrain the SR_θ network to generate images with the same physical acquisition properties as the initial I^{LR} images.

3.4. Super-resolution network

We decided to use the layout for the SR network proposed in Ledig et al. (2017). SR_θ is aimed at producing images that are similar to the one in the target domain by trying to fool the discriminator network. This is achieved through the term l_{Adv} in the proposed loss function defined as follows:

$$l_{Adv} = -logDS_\gamma(SR_\theta(I^{LR})), \qquad (5)$$

where $DS_\gamma(SR_\theta(I^{LR}))$ is the probability that the predicted image $SR_\theta(I^{LR})$ is classified as a real T^{HR}. As proposed by Goodfellow et al. (2014) we minimize $-logDS_\gamma(SR_\theta(I^{LR}))$ instead of $log[1 - DS_\gamma(SR_\theta(I^{LR}))]$ for better gradient behaviour.

In the inference phase, only SR_θ is used for processing the I^{LR} images.

3.5. Regularization

The blocks displayed in orange in Fig. 2 are used to regularize the network training. This regularization is required since the Voronoi vectorization of each patch is normalized to the range [0–1] and this may result in an expansion of its histogram range. To restore the correct histogram distribution, we impose that the mean values in each row and each column of the patch are identical between the initial I^{LR} and the obtained T^{HR}. This is achieved in our framework through the l_{Reg} term of $loss_t$:

$$l_{Reg} = \frac{1}{H}\sum_{y=1}^{H}\left[\frac{1}{W}\sum_{x=1}^{W}SR_\theta(I_{xy}^{LR}) - \frac{1}{W}\sum_{x=1}^{W}I_{xy}^{LR}\right]^2$$
$$+ \frac{1}{W}\sum_{x=1}^{W}\left[\frac{1}{H}\sum_{y=1}^{H}SR_\theta(I_{xy}^{LR}) - \frac{1}{H}\sum_{y=1}^{H}I_{xy}^{LR}\right]^2 \qquad (6)$$

3.6. Training domain

In our pipeline we considered four different target domains to transfer the super-resolution to the initial LR images: i) T_{nat}^{HR} where the HR patches are extracted from natural images (grey-scaled images from the Sun2012 database (Xiao et al., 2010)), ii) T_{orig}^{HR} containing the HR patches obtained by the video-registration technique on the LR images, iii) T_{syn}^{HR} containing paired HR patches obtained using the video-registration technique while the LR are synthetically aligned, and iv) T_{res}^{HR} where the HR patches are obtained by down-sampling large portions of the LR images by a factor of four. Inspired by the work proposed in Shocher et al. (2017),

the idea behind this last target domain is based on the fact that patches in the images have recurrences at a different scale and down-sampling large LR images may increase the high-frequency responses in the generated down-sampled HR patches.

3.7. Training details and parameters

In our implementation, Eq. (1) is solved by minimizing the cross-entropy of the number of samples correctly discriminated by *DS*. As proposed by Arjovsky and Bottou (2017) we add white noise to the inputs of the DS_γ network to stabilize the adversarial training. We trained our networks on an NVIDIA GTX TITAN-X GPU card with 12GB of memory. The training procedure converges after 50–80 thousand iterations of random mini-batch with 54 patches. For the optimization of the stochastic gradient descent, we use Adam with $\beta1 = 0.9$, $\beta2 = 0.999$ and $\epsilon = 10e-8$. The networks were trained with a learning rate of $10e-4$.

4. Experiments

Due to the lack of real ground truth in our database, the validation of our experiments is based on complementary quantitative and qualitative analysis. The quantitative analysis, presented in Section 4.1, uses four different metrics to evaluate the obtained images. The qualitative analysis is instead based on a MOS study carried out by clinicians and medical imaging experts that gave numerical indications of the perceived quality of the super-resolved images.

4.1. Quantitative analysis

The four metrics used in our quantitative analysis are: i) a Structural Similarity matrix (SSIM) proposed by Wang et al. (2004) that evaluates the similarity between $SR_\theta(I^{LR})$ and $\widehat{HR}$, ii) $\triangle GCF_{\widehat{HR}}$ that quantifies the improvement on the global contrast factor (a reference-free metric for measuring image contrast Matkovic et al., 2005) that the super-resolved image yields with respect to $\widehat{HR}$, iii) $\triangle GCF_{ILR}$ that is the improvement of the global contrast factor that the super-resolved image yields with respect to the initial I^{LR}, and iv) a composite score Tot_{cs} obtained by normalizing the value of SSIM and $\triangle GCF_{\widehat{HR}}$ in the range [0,1] and averaging the obtained results. The formula used to compute Tot_{cs} is described by Eq. (7).

$$Tot_{cs} = \frac{\frac{SSIM_{\widehat{HR}}-0.6}{0.4} + \frac{\triangle GCF_{\widehat{HR}}+0.5}{1.8}}{2} \qquad (7)$$

This composite score leads to a more robust evaluation of the results since, SSIM alone is not reliable when the ground truth is only estimated, while the GCF can be improved by merely adding random high frequency to the images.

Table 1

Quantitative analysis results obtained by our approach when trained with different target domains on case study CS_1 and CS_2.

		T_{syn}^{HR}	T_{orig}^{HR}	T_{res}^{HR}	T_{nat}^{HR}
CS_1	$SSIM_{\widehat{HR}}$	0.90 ± 0.03	$\mathbf{0.91 \pm 0.03}$	0.87 ± 0.04	0.86 ± 0.03
	$\triangle GCF_{\widehat{HR}}$	0.01 ± 0.29	0.38 ± 0.27	-0.13 ± 0.40	$\mathbf{0.66 \pm 0.31}$
	$\triangle GCF_{lR}$	-0.28 ± 0.19	0.09 ± 0.19	-0.42 ± 0.31	$\mathbf{0.38 \pm 0.26}$
	Tot_{cs}	0.52	0.63	0.44	$\mathbf{0.64}$
CS_2	$SSIM_{\widehat{HR}}$	$\mathbf{0.91 \pm 0.03}$	$\mathbf{0.91 \pm 0.03}$	0.87 ± 0.03	0.86 ± 0.04
	$\triangle GCF_{\widehat{HR}}$	-0.10 ± 0.36	0.24 ± 0.35	-0.26 ± 0.40	$\mathbf{0.51 \pm 0.40}$
	$\triangle GCF_{lR}$	-0.33 ± 0.21	0.01 ± 0.18	-0.49 ± 0.28	$\mathbf{0.28 \pm 0.25}$
	Tot_{cs}	0.49	0.59	0.41	$\mathbf{0.61}$

Table 2

Quantitative analysis results of the proposed approach against state-of-the-art methods on the database DB_{orig} for case study CS_1 and CS_2.

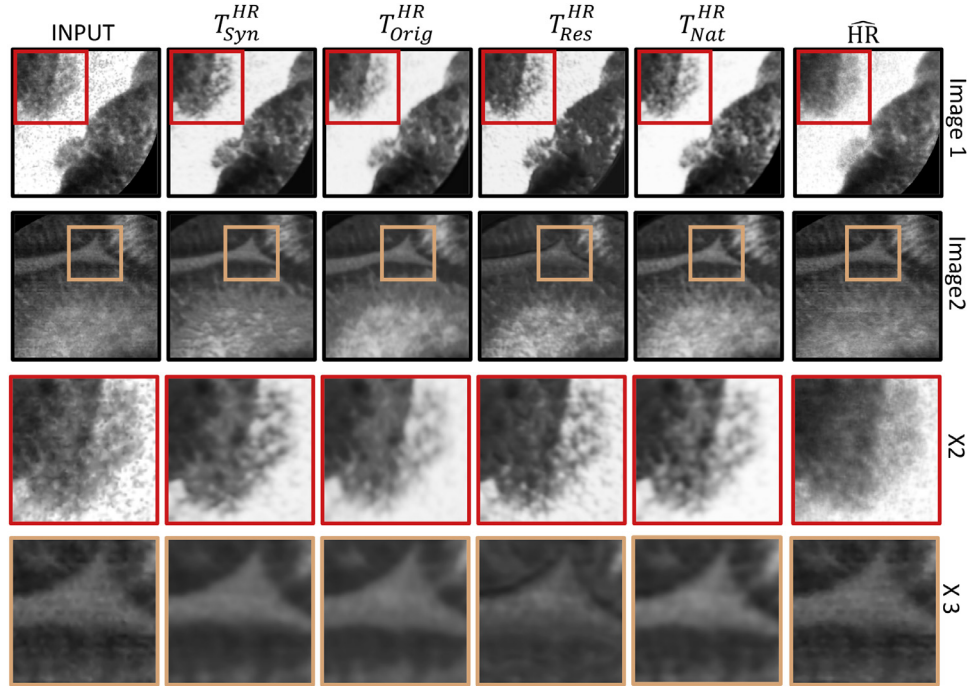|		Proposed	Ravì et al. (2018)	Villena et al. (2013)	Wiener	Contrast-enhancement
CS_1	$SSIM_{\widehat{HR}}$	0.86 ± 0.03	$\mathbf{0.88 \pm 0.05}$	0.86 ± 0.05	0.83 ± 0.08	0.62 ± 0.08
	$\triangle GCF_{\widehat{HR}}$	0.66 ± 0.31	0.42 ± 0.24	0.27 ± 0.23	-0.00 ± 0.37	$\mathbf{1.34 \pm 0.36}$
	$\triangle GCF_{lR}$	0.38 ± 0.26	0.13 ± 0.13	-0.02 ± 0.07	-0.29 ± 0.27	$\mathbf{1.06 \pm 0.25}$
	Tot_{cs}	$\mathbf{0.64}$	0.61	0.54	0.42	0.53
CS_2	$SSIM_{\widehat{HR}}$	0.86 ± 0.04	$\mathbf{0.89 \pm 0.04}$	0.88 ± 0.04	0.85 ± 0.06	0.63 ± 0.06
	$\triangle GCF_{\widehat{HR}}$	0.51 ± 0.40	0.38 ± 0.29	0.21 ± 0.29	-0.15 ± 0.44	$\mathbf{1.32 \pm 0.34}$
	$\triangle GCF_{lR}$	0.28 ± 0.25	0.15 ± 0.09	-0.02 ± 0.09	-0.38 ± 0.31	$\mathbf{1.09 \pm 0.18}$
	Tot_{cs}	$\mathbf{0.61}$	0.60	0.54	0.41	0.54

Fig. 5. Example of visual results obtained by the proposed approaches when trained with different target domains. From left to right we have: Input, training with T_{syn}^{HR}, training with T_{orig}^{HR}, training with T_{res}^{HR}, training with T_{nat}^{HR} and $\widehat{HR}$.

Our first experiment is aimed at finding the best target domain for improving the pCLE images. The results computed on DB_{orig} for both the case studies are reported in Table 1. As we can see, the network trained with natural images (T_{nat}^{HR}) obtains the best Tot_{cs} score. From these results, we can also deduct that using synthetic images for the training is worse than using images from the original domain. This is probably due to the fact that synthetic images may have a non-negligible domain gap with the real images. With this result, we can state that paired patches are not anymore a requirement for our framework. Finally, downsampling LR images to create patches with higher frequency content does not seem to provide good results and the high-frequency signals are not recov-

ered. These qualitative indications can be seen on reconstructed images reported in Fig. 5.

Looking at the different case studies, the aforementioned considerations are consistent along both the cases, although CS_2 shows slightly lower performances with respect to CS_1 probably due to the fact that it has a coarser split of its dataset.

To further validate our framework, we compare our best approach (the network trained with the target domain T_{nat}^{HR}), against some state-of-the-art single image super-resolution methodologies. These results are presented in Table 2 for the database DB_{orig} and in Table 3 for the database DB_{syn}. In these experiments we consider three different approaches: i) the unsupervised

Table 3

Quantitative analysis results of the proposed approach against state-of-the-art methods on the database DB_{syn} for case study CS_1 and CS_2.

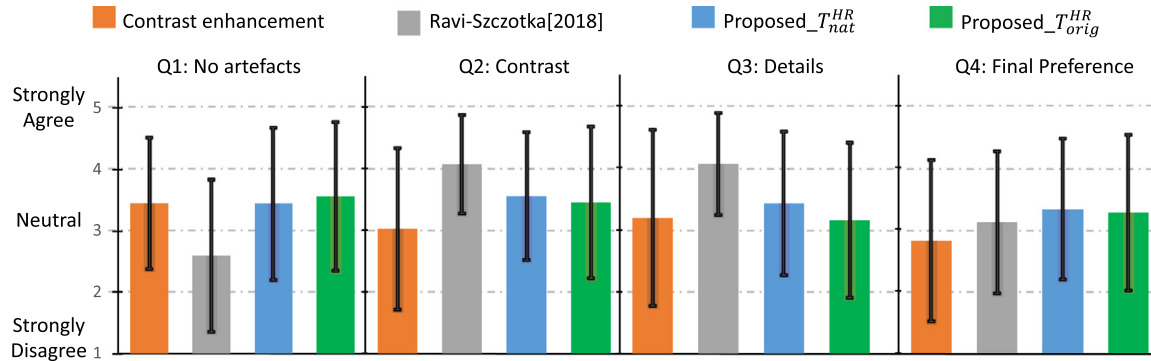|		Proposed	Ravì et al. (2018)	Villena et al. (2013)	Wiener	Contrast-enhancement
CS_1	$SSIM_{\widehat{HR}}$	0.90 ± 0.03	$\mathbf{0.93 \pm 0.03}$	0.89 ± 0.05	0.88 ± 0.06	0.66 ± 0.08
	$\triangle GCF_{\widehat{HR}}$	0.60 ± 0.22	0.45 ± 0.25	0.13 ± 0.11	-0.21 ± 0.21	$\mathbf{1.03 \pm 0.32}$
	$\triangle GCF_{\mathit{LR}}$	0.47 ± 0.26	0.32 ± 0.32	0.00 ± 0.00	-0.34 ± 0.19	$\mathbf{0.90 \pm 0.26}$
	Tot_{cs}	$\mathbf{0.68}$	$\mathbf{0.68}$	0.54	0.43	0.50
CS_2	$SSIM_{\widehat{HR}}$	0.91 ± 0.03	$\mathbf{v0.92 \pm 0.03}$	0.90 ± 0.04	0.89 ± 0.06	0.65 ± 0.07
	$\triangle GCF_{\widehat{HR}}$	0.50 ± 0.25	0.52 ± 0.20	0.11 ± 0.05	-0.28 ± 0.22	$\mathbf{1.08 \pm 0.20}$
	$\triangle GCF_{\mathit{LR}}$	0.39 ± 0.23	0.41 ± 0.23	-0.00 ± 0.00	-0.38 ± 0.21	$\mathbf{0.97 \pm 0.18}$
	Tot_{cs}	0.66	$\mathbf{0.68}$	0.54	0.42	0.50

Fig. 6. Mean and standard deviation of the participants' replies to each of the four MOS questions for the evaluation of the results obtained by contrast-enhancement (baseline), Ravì et al. (2018), and proposed approach trained using two different target domains (T_{nat}^{HR} and T_{orig}^{HR}). These results were obtained on the test images of DB_{orig} for the case study CS_1.

Wiener deconvolution tuned on the train set, ii) the unsupervised variational Bayesian inference approach with sparse and non-sparse priors (Villena et al., 2013), and the supervised EBSR proposed by Ravì et al. (2018). Finally, a contrast-enhancement approach obtained by sharpening the input was also used as a baseline.

In the dataset DB_{orig}, although the sharpening algorithm produces the best contrast improvements, our approach obtains the highest SSIM and, according to Tot_{cs}, the overall performance outperforms all the other approaches.

Differently from the results obtained with DB_{orig}, with the database DB_{syn}, our approach is not able to overcome the results obtained by Ravì et al. (2018). This is probably due to the fact that the supervised training in Ravì et al. (2018) exploits principles that are similar to the ones used to generate the synthetic images in DB_{syn}. Therefore, the results obtained by this approach in the database DB_{syn} are obtained in a purely simulated scenario where the model is trained on data with no domain gap with the test set. Consequently, the supervised approach in Ravì et al. (2018) has an advantage with respect to our unsupervised one in this setting. What is interesting to see nonetheless, is that our unsupervised framework can achieve almost similar performance to the supervised one of (Ravì et al., 2018) despite the evaluation being intrinsically favourable for this last solution.

Also in these two experiments, close results are obtained between CS_1 and CS_2 confirming that the system is able to super-resolve images for both the considered cases (i.e. when the images contain new visual structures or when they are extracted from new patients).

The statistical significances of the improvements discussed in this section were assessed with a paired t-test and the p-values are all less than 0.0001.

4.2. Semi-quantitative analysis (MOS)

To perform the MOS, we asked 10 trained individuals to evaluate, on average, 20 images each, randomly selected from the test set of DB_{orig} on the case study CS_1. At each step, the SR images obtained with two different configurations of the proposed approach, with Ravì et al. (2018), and with a contrast-enhancement approach that sharpens the input (baseline), were shown to the user in a random order to reduce any possible bias on the evaluation of the images. The two configurations used for our approach were the one based on training our model with the two best target domains (i.e. T_{nat}^{HR} and T_{orig}^{HR}). The input and the $\widehat{HR}$ were also displayed on the screen as references for the participants. For each of the four images, the user assigned a score between 1 (strongly disagree) to 5 (strongly agree) on the following questions:

- Q1: Is the image artefact-free?
- Q2: Can you see an improvement in contrast with respect to the input?
- Q3: Can you see an improvement in the details with respect to the input?
- Q4: Would you prefer seeing the new image over the input?

To make sure that the questions were consistently interpreted, each participant received a short training before starting the study. The results on the MOS presented in Fig. 6 show that, between the two different configurations used on our approach, the model trained with natural images T_{nat}^{HR} provides a better trade-off of artefacts, contrast and details with respect to the training using the target domain T_{orig}^{HR}. The results show also that the proposed approach and Ravì et al. (2018) provide complementary features. In fact, although the details (question Q3) and the contrast (question Q2) in both the settings of our approach seem to be worse than Ravì et al. (2018), our solution provides better scores for the absence of artefacts (question Q1), which is an important characteristic in clinical applications. Regarding the final preference (question Q4), our solution trained using natural images (T_{nat}^{HR}) shows the best results with respect to all the other approaches, confirming the validity of our solution to perform super-resolution on pCLE images. The approach that sharpens the images is, instead,
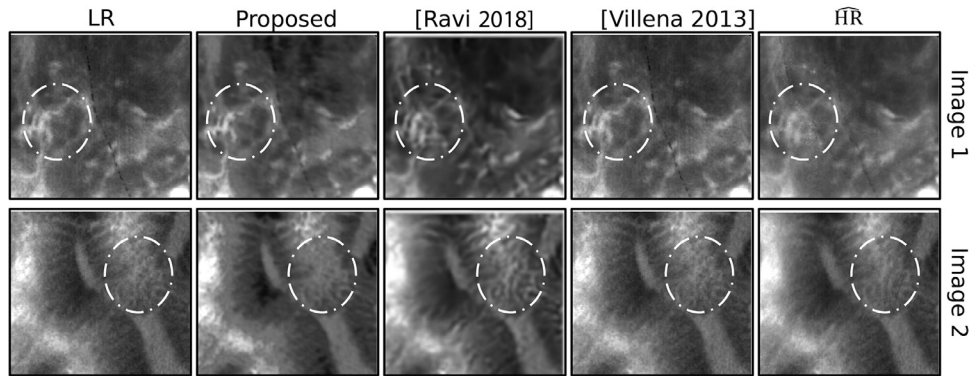
Fig. 7. Example of visual results obtained by our approach in comparison with other state-of-the-art approaches. From left to right we have: Input image, proposed output, output from Ravì et al. (2018), output from Villena et al. (2013) and $\widehat{HR}$.

the one that provides the lowest scores for Q2, Q3 and Q4, probably because it enhances the noise.

Visual results for some of these images are shown in Fig. 7, confirming these findings. More specifically, although the output proposed by Ravì et al. (2018) shows better contrast and higher sharpness images, the corresponding algorithm also behaves much worse in term of artefacts generation. In fact, as we can see in Fig. 7, it often enhances noise and makes up details that are not visible neither in the input nor in the estimated HR images and this can eventually lead to a wrong clinical interpretation of the images. In Fig. 7 we have marked with a white circle some of the regions where these issues are more evident. From a clinical point of view, we believe that the reliability of SR images in terms of details is a more important feature than having high contrast or high sharpness created by artefacts. According to these visual considerations and the MOS findings, we can conclude that our solution provides a more convincing representation for super-resolved pCLE images with respect to the other state-of-the-art approaches.

5. Discussion and conclusions

Obtaining medical images that accurately visualize structures of tissues is still today an open challenge. One of the main issue that researchers are trying to address here, is to improve the image resolution. In endomicroscopy, low image resolution is often dependent on the intrinsic limitations of the acquisition systems. Current solutions propose SR methods to post-process the final images as an alternative to the more difficult hardware enhancements. Clinical impacts and benefits in the use of SR methods include: i) better localization of tissue structures, ii) improving the image contrast and iii) improving the Signal to Noise Ratio (SNR) (Greenspan, 2008). However, often the validation of these benefits in terms of clinical outcome is not straightforward. An attempt for this was proposed in Kennedy et al. (2006), where Positron Emission Tomography (PET) scans on phantom and patients were used to prove that smaller visual features were localized and better visualized using SR techniques than without. Another similar study conducted in Kennedy et al. (2007) shows that using SR techniques produces better contrast ratios and better target-to-background ratios than the standard reconstructions. Plenge et al. (2012) designed, instead, an experimental framework to show that the SR reconstructions are more advantageous in terms of the SNR with respect to the direct HR acquisition. Finally, Sano et al. (2017) proposed a novel measurement algorithm for joint space distance on X-ray images generated by a SR method. The results exhibit higher accuracy in the measured distances when SR images were used.

The studies above show that SR methods can improve the clinical outcomes and can open the door for better diagnosis. In our case, however, the lack of a real ground truth can raise some scepticism on the validation of the results since is not simple to show that SR approaches don't emphasize or make up details that are not real. Our extensive quantitative and qualitative analysis, based also on expert's evaluations, are developed to show the reliability of the obtained SR images and the reduced presence of artefacts even in the absence of real ground truth. More specifically, these results validated using two versions of a database containing 238 endomicroscopy video sequences captured from 143 patients demonstrate the ability of the pipeline to produce convincing super-resolved images.

In conclusion, in our study, we report a super-resolution framework for endomicroscopy images based on an unsupervised adversarial deep neural network that takes advantage of the knowledge of the physical acquisition process to impose a cycle consistency. The proposed framework results to be particularly useful in all situations where there is a lack of HR images and pairs of LR/HR images are not available for the supervised training.

To the best of our knowledge, we are the first to propose an unsupervised super-resolution approach for medical images. Further clinical studies could validate the relevance of the proposed framework to specific clinical applications for super-resolution.

References

André, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N., 2011. A smart atlas for endomicroscopy using automated video retrieval. Med. Image Anal. 15 (4), 460–476.

Arjovsky, M., Bottou, L., 2017. Towards principled methods for training generative adversarial networks. Stat 1050, 17.

Ayasso, H., Rodet, T., Abergel, A., 2012. A variational bayesian approach for unsupervised super-resolution using mixture models of point and smooth sources applied to astrophysical map-making. Inverse Probl. 28 (12), 125005.

Bao, J., Chen, D., Wen, F., Li, H., Hua, G., 2017. Cvae-gan: fine-grained image generation through asymmetric training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2745–2754.

Bulat, A., Yang, J., Tzimiropoulos, G., 2018. To learn image super-resolution, use a GAN to learn how to do image degradation first. arXiv:1807.11458.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680.

Greenspan, H., 2008. Super-resolution in medical imaging. Comput. J. 52 (1), 43–63.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision. Springer, pp. 694–711.

Kennedy, J.A., Israel, O., Frenkel, A., Bar-Shalom, R., Azhari, H., 2006. Super-resolution in pet imaging. IEEE Trans. Med. Imaging 25 (2), 137–147.

Kennedy, J.A., Israel, O., Frenkel, A., Bar-Shalom, R., Azhari, H., 2007. Improved image fusion in pet/ct using hybrid image reconstruction and super-resolution. Int. J. Biomed. Imaging 2007.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690.

Matkovic, K., Neumann, L., Neumann, A., Psik, T., Purgathofer, W., 2005. Global contrast factor - a new approach to image contrast.. Comput. Aesthetics 2005, 159–168.

Mescheder, L., Nowozin, S., Geiger, A., 2017. Adversarial variational bayes: unifying variational autoencoders and generative adversarial networks. In: International Conference on Machine Learning, pp. 2391–2400.

Neumann, H., Kiesslich, R., Wallace, M.B., Neurath, M.F., 2010. Confocal laser endomicroscopy: technical advances and clinical applications. Gastroenterology 139 (2), 388–392.

Nguyen, D.L., Lee, J.G., Parekh, N.K., Samarasena, J., Bechtold, M.L., Chang, K., 2015. The current and future role of endomicroscopy in the management of inflammatory bowel disease. Ann. Gastroenterol. 28 (3), 331.

Plenge, E., Poot, D.H., Bernsen, M., Kotek, G., Houston, G., Wielopolski, P., van der Weerd, L., Niessen, W.J., Meijering, E., 2012. Super-resolution reconstruction in MRI: better images faster? In: Medical Imaging 2012: Image Processing, 8314. International Society for Optics and Photonics, p. 83143V.

Ravì, D., Fabelo, H., Callic, G.M., Yang, G.-Z., 2017. Manifold embedding and semantic segmentation for intraoperative guidance with hyperspectral brain imaging. IEEE Trans. Med. Imaging 36 (9), 1845–1857.

Ravì, D., Szczotka, A.B., Shakir, D.I., Pereira, S.P., Vercauteren, T., 2018. Effective deep learning training for single-image super-resolution in endomicroscopy exploiting video-registration-based reconstruction. Int. J. Comput. Assist. Radiol. Surg. 13 (6), 917–924.

Sano, Y., Mori, T., Goto, T., Hirano, S., Funahashi, K., 2017. Super-resolution method and its application to medical image processing. In: Consumer Electronics (GCCE), 2017 IEEE 6th Global Conference on. IEEE, pp. 1–2.

Shocher, A., Cohen, N., Irani, M., 2017. "zero-shot" super-resolution using deep internal learning. arXiv:1712.06087.

Torre, L.A., Siegel, R.L., Ward, E.M., Jemal, A., 2016. Global cancer incidence and mortality rates and trends-an update. Cancer Epidemiol. Prev. Biomark. 25 (1), 16–27.

Villena, S., Vega, M., Babacan, S.D., Molina, R., Katsaggelos, A.K., 2013. Bayesian combination of sparse and non-sparse priors in image super resolution. Digit. Signal Process. 23 (2), 530–541.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A., 2010. Sun database: large-scale scene recognition from abbey to zoo. In: Computer Vision and Pattern Recognition, 2010 IEEE Conference on. IEEE, pp. 3485–3492.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2223–2232.