OPTIMAL TRANSPORT, CYCLEGAN, AND PENALIZED LS FOR UNSUPERVISED LEARNING IN INVERSE PROBLEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

The penalized least squares (PLS) is a classic approach to inverse problems, where a regularization term is added to stabilize the solution. Optimal transport (OT) is another mathematical framework for computer vision tasks by providing means to transport one measure to another at minimal cost. Cycle-consistent generative adversarial network (cycleGAN) is a recent extension of GAN to learn target distributions with less mode collapsing behavior. Although similar in that no supervised training is required, the algorithms look different, so the mathematical relationship between these approaches is not clear. In this article, we provide an important advance to unveil the missing link. Specifically, we reveal that a cycleGAN architecture can be derived as a dual formulation of the optimal transport problem, if the PLS with a deep learning penalty is used as a transport cost between the two probability measures from measurements and unknown images. This suggests that cycleGAN can be considered as stochastic generalization of classical PLS approaches. Our derivation is so general that various types of cycleGAN architecture can be easily derived by merely changing the transport cost. As proofs of concept, this paper provides novel cycleGAN architecture for unsupervised learning in accelerated MRI and deconvolution microscopy problems, which confirm the efficacy and the flexibility of the theory.

1 INTRODUCTION

Inverse problems are ubiquitous in computer vision, biomedical imaging, scientific discovery, etc. In inverse problems, a noisy measurement $y \in \mathcal{Y}$ from an unobserved image $x \in \mathcal{X}$ is modeled by

$$y = \mathcal{H}x + w,$$ (1)

where w is the measurement noise, and $\mathcal{H} : \mathcal{X} \mapsto \mathcal{Y}$ is the measurement operator. In inverse problems originated from physics, the measurement operator is usually represented by an integral equation:

$$\mathcal{H}x(\boldsymbol{r}) := \int_{\mathbb{R}^d} h(\boldsymbol{r}, \boldsymbol{r}')x(\boldsymbol{r}')d\boldsymbol{r}', \quad \boldsymbol{r} \in \mathcal{D} \subset \mathbb{R}^d.$$ (2)

for $d = 2, 3$, where $h(\boldsymbol{r}, \boldsymbol{r}'; x)$ is an integral kernel. Then, the inverse problem is formulated as an estimation problem of the unknown x from the measurement y. It is well-known that an inverse problem is ill-posed. A classical strategy to mitigate the ill-posedness is the penalized least squares (PLS) approach:

$$\hat{x} = \arg\min_x c(x; y) := \|y - \mathcal{H}x\|^q + R(x)$$ (3)

for $q \geq 1$, where $R(x)$ is a regularization (or penalty) function such as l_1, total variation (TV), etc (Chaudhuri et al., 2014; Sarder & Nehorai, 2006; McNally et al., 1999). In some inverse problems, the measurement operator $\mathcal{H}$ is not well defined so that both the unknown operator $\mathcal{H}$ and the image x should be estimated.

Recently, deep learning approaches with supervised training have become the main approaches to inverse problems because of their excellent and ultra-fast reconstruction performance. For example,

in the low-dose x-ray computed tomography (CT) denoising problems, a convolutional neural network (CNN) is trained to learn the relationship between noisy image y and the matched noiseless (or high dose) label images x (Kang et al., 2017). In the context of (3), the supervised neural network can be understood as directly learning the operation $\hat{x} = \arg\min_x c(x; y)$. Unfortunately, in many applications, matched label data are not available for supervised training. Therefore, unsupervised training without matched reference data has become an important research topic.

Recently, generative adversarial network (GAN) has attracted significant attention in machine learning community by providing a way to generate target data distribution from random distribution (Goodfellow et al., 2014). In particular, the authors in (Arjovsky et al., 2017) proposed so-called Wasserstein GAN (W-GAN), which is closely related to the mathematical theory of optimal transport (OT) (Villani, 2008; Peyré et al., 2019). In optimal transport, for given two probability measures supported on the $\mathcal{X}$ and $\mathcal{Y}$ spaces, one pays a cost for transporting one measure to another. Then, the minimization of the average transportation cost provides an unsupervised way of learning the transport map between the two measures. Unfortunately, these GAN approaches often generate artificial features due to the mode collapsing, so cycleGAN (Zhu et al., 2017) that imposes the one-to-one correspondency has been extensively investigated (Kang et al., 2019; Lu et al., 2017).

Although classical PLS, optimal transport, and cycleGAN share the commonality of unsupervised learning which does not require matched training data, there is no mathematical theory to systematically link these seemingly different approaches. Therefore, one of the main contributions of this paper is to unveil the missing link between these methods. In particular, we reveal that cycleGAN architecture can be derived as a dual formulation of the optimal transport problem when the cost function of PLS with a deep learning penalty is used as a transport cost between the two measures. Thus, cycleGAN can be considered as stochastic generalization of PLS. Moreover, this framework is so general that various cycleGAN architecture can be easily obtained by changing the PLS cost.

As a proof of concept, we provide novel cycleGAN architecture for unsupervised learning in two physical inverse problems: accelerated MRI, and deconvolution microscopy. In these applications, we show that only one single CNN generator is required, which significantly reduces the training complexity. Experimental results confirmed that the proposed unsupervised learning approaches can successfully provide accurate inversion results without any matched reference.

2 RELATED WORKS

2.1 OPTIMAL TRANSPORT

Optimal transport compares two measures in a Lagrangian framework (Villani, 2008; Peyré et al., 2019). Formally, we say that $T : \mathcal{X} \mapsto \mathcal{Y}$ transports $\mu \in P(\mathcal{X})$ to $\nu \in P(\mathcal{Y})$, if

$$\nu(B) = \mu\left(T^{-1}(B)\right), \quad \text{for all } \nu\text{-measureable set } B, \tag{4}$$

which condition is often simply represented by $\nu = T_{\#}\mu$, where $T_{\#}$ is often called the push-forward operator. Monge's original optimal transport problem (Villani, 2008; Peyré et al., 2019) is then to find a transport map T that transport μ to ν at the minimum total transportation cost:

$$\min_{T} \quad \mathbb{M}(T) \quad := \int_{\mathcal{X}} c(x, T(x))d\mu(x) \tag{5}$$
$$\text{subject to} \quad \nu = T_{\#}\mu$$

However, this is usually computational expensive due to the nature of combinatorial assignment. Kantorovich relaxed the assumption to consider a probabilistic transport, which allows mass splitting from a source toward several targets (Villani, 2008; Peyré et al., 2019). Specifically, Kantorovich introduced a joint measure $\pi \in P(\mathcal{X} \times \mathcal{Y})$ and the associated cost $c(x, y), x \in \mathcal{X}, y \in \mathcal{Y}$ such that $d\pi(x, y)$ becomes the amount of mass transferred from x to y and $c(x, y)d\pi(x, y)$ is the associated cost. Then, the Kantorovich's relaxation is formulated as

$$\min_{\pi} \quad \mathbb{K}(\pi) \quad := \int_{\mathcal{X} \times \mathcal{Y}} c(x, y)d\pi(x, y) \tag{6}$$

$$\text{subject to} \quad \pi(A \times \mathcal{Y}) = \mu(A), \ \pi(\mathcal{X} \times B) = \nu(B)$$

for all measurable set $A \in \mathcal{X}$ and $B \in \mathcal{Y}$. Here, the last two constraints come from the observation that the total amount of mass removed from any measurable set has to equal to the marginals (Villani, 2008; Peyré et al., 2019). Another important advantage of Kantorovich formulation is the dual formulation as stated in the following theorem:

Theorem 1 (Kantorovich duality theorem). *(Villani, 2008, Theome 5.10, p.57-p.59) Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two Polish probability spaces (separable complete metric space) and let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous cost function, such that $|c(x,y)| \leq c_{\mathcal{X}}(x) + c_{\mathcal{Y}}(y)$ for some $c_{\mathcal{X}} \in L^1(\mu)$ and $c_{\mathcal{Y}} \in L^1(\nu)$, where $L^1(\mu)$ denotes the set of 1-Lipschitz functions with the measure μ. Then, there is duality:*

$$\min_{\pi \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x,y) d\pi(x,y) = \max_{\varphi \in L^1(\mu)} \left\{ \int_{\mathcal{X}} \varphi(x) d\mu(x) + \int_{\mathcal{Y}} \varphi^c(y) d\nu(y) \right\}$$

and the above maximum is taken over so-called Kantorovich potential φ *whose c-transform* $\varphi^c(y) = \sup_x(c(x,y) - \varphi(x))$ *is properly defined.*

2.2 PENALIZED LEAST SQUARE WITH DEEP LEARNING PRIOR

Recently, model-based image reconstruction frameworks using a convolution neural network (CNN) based prior have been extensively studied (Zhang et al., 2017; Aggarwal et al., 2018) thanks to their similarities to the classical regularization theory. The main idea is to utilize a pre-trained neural network to stabilize the inverse solution. Specifically, the problem can be formulated as

$$\min_x c(x; y, \Theta, \mathcal{H}) = \|y - \mathcal{H}x\|^q + \|G_\Theta(y) - x\|^p \tag{7}$$

for $p \geq 1$, where $G_\Theta(y)$ is a pre-trained CNN with the network parameter Θ and the input y. The minimization problem of (7) is to find a balance between data fidelity term and the CNN output.

3 MAIN CONTRIBUTIONS

One of the basic assumptions in the classical PLS formulation is that the measurement y is fixed and one is interested in finding unknown x assuming some prior distribution. In this section, we relax the assumption and consider unsupervised learning situations where both of them are random samples from joint probability measure $\pi(x,y)$. Interesting, this leads to cycleGAN architecture. Specifically, by considering x and y in (7) as random variables, we have the following reformulation of the cost:

$$c(x, y; \Theta, \mathcal{H}) = \|y - \mathcal{H}x\|^q + \|G_\Theta(y) - x\|^p \tag{8}$$

where the first two terms before the semi-colon are the random vectors. Unlike (7), the network $G_\Theta(y)$ in (8) is not pretrained and needs to be learned. More specifically, $\Theta, \mathcal{H}$ are the network and measurement system parameters that should be estimated. These parameters can be found by minimizing the average transport cost for all combinations of $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ with respect to the joint measure $\pi(x,y)$:

$$\mathbb{T}(\Theta, \mathcal{H}) := \min_\pi \int_{\mathcal{X} \times \mathcal{Y}} c(x, y; \Theta, \mathcal{H}) d\pi(x,y) \tag{9}$$

where the minimum is taken over all joint distribution whose marginal distribution with respect to X and Y is μ and ν, respectively. Below we show that the average transportation cost in (9) has an interesting decomposition.

Lemma 1. *If the mapping $G_\Theta : \mathcal{Y} \mapsto \mathcal{X}$ is single-valued, then the average transportation cost $\mathbb{T}(\Theta, \mathcal{H})$ in (9) can be decomposed as*

$$\mathbb{T}(\Theta, \mathcal{H}) = \ell_{cycle}(\Theta, \mathcal{H}) + \ell_{OT'}(\Theta, \mathcal{H}) \tag{10}$$

where

$$\ell_{cycle}(\Theta, \mathcal{H}) = \min_\pi \int_{A \cup B} c(x, y; \Theta, \mathcal{H}) d\pi(x,y) \tag{11}$$

$$\ell_{OT'}(\Theta, \mathcal{H}) = \min_\pi \int_{\mathcal{X} \times \mathcal{Y} \setminus A \cup B} c(x, y; \Theta, \mathcal{H}) d\pi(x,y) \tag{12}$$

and

$$A = \{(x,y) \in \mathcal{X} \times \mathcal{Y} \mid x = G_\Theta(y),\ y \in \mathcal{Y}\}, \quad B = \{(x,y) \in \mathcal{X} \times \mathcal{Y} \mid y = \mathcal{H}x,\ x \in \mathcal{X}\}$$

Proof. From the definition of ℓ_{cycle} and $\ell_{OT'}$, it is straightforward to show that

$$\mathbb{T}(\Theta, \mathcal{H}) \geq \ell_{cycle}(\Theta, \mathcal{H}) + \ell_{OT'}(\Theta, \mathcal{H}).$$

To show reversed inequality, we need the classical results on optimal transport for the restricted measure (Villani, 2008). Specifically, Theorem 2 in Appendix informs that for some optimal transportation plan π^*, the restrictions $\pi^*|_{A \cup B}$ and $\pi^*|_{\mathcal{X} \times \mathcal{Y} \setminus (A \cup B)}$ acquire optimality. Therefore,

$$
\begin{aligned}
\mathbb{T}(\Theta, \mathcal{H}) &= \int_{\mathcal{X} \times \mathcal{Y}} c(x, y; \Theta, \mathcal{H}) d\pi^*(x, y) \\
&= \int_{A \cup B} c(x, y; \Theta, \mathcal{H}) d\pi^*(x, y) + \int_{\mathcal{X} \times \mathcal{Y} \setminus (A \cup B)} c(x, y; \Theta, \mathcal{H}) d\pi^*(x, y) \\
&= \ell_{cycle}(\Theta, \mathcal{H}) + \ell_{OT'}(\Theta, \mathcal{H})
\end{aligned}
$$

This concludes the proof. $\square$

The following proposition tells why we name the first term as ℓ_{cycle}. Indeed, the term $\ell_{cycle}(\Theta, \mathcal{H})$ in (11) is the weighted version of cycle consistency term in cycleGAN (Zhu et al., 2017), which imposes the constraint $x \simeq G_\Theta(\mathcal{H}x), \forall x \in \mathcal{X}$ and $y \simeq \mathcal{H}G_\Theta(y), \forall y \in \mathcal{Y}$. However, our formulation is more general, since there exists weighting factors that can balance the importance of the cycle-consistency terms.

Proposition 1.

$$
\begin{aligned}
\ell_{cycle}(\Theta, \mathcal{H}) &:= \min_{\pi} \int_{\mathcal{X} \times \mathcal{Y} \setminus A \cup B} c(x, y; \Theta, \mathcal{H}) d\pi(x, y) \\
&= \int_{A \cup B} c(x, y; \Theta, \mathcal{H}) \pi^*(dx, dy) \\
&= \int_{\mathcal{X}} \|x - G_\Theta(\mathcal{H}x)\|^p \rho(x) \mu(dx) + \int_{\mathcal{Y}} \|y - \mathcal{H}G_\Theta(y)\|^q \sigma(y) \nu(dy) \quad (13)
\end{aligned}
$$

where ρ and σ are measurable functions such that $0 \leq \rho(x) \leq 1$ μ-a.e., and $0 \leq \sigma(y) \leq 1$ ν-a.e.

Proof. By virtue of the definitions of the cost function and the sets A and B, we have

$$
\begin{aligned}
c(x, y; \Theta, \mathcal{H}) &= \|y - \mathcal{H}G_\Theta(y)\|^q \text{ on } A, \\
c(x, y; \Theta, \mathcal{H}) &= \|x - G_\Theta(\mathcal{H}x)\|^p \text{ on } B, \\
c(x, y; \Theta, \mathcal{H}) &= 0 \qquad \text{on } A \cap B.
\end{aligned}
$$

Hence,

$$\int_{A \cup B} c(x, y; \Theta, \mathcal{H}) d\pi^*(x, y) = \int_B \|x - G_\Theta(\mathcal{H}x)\|^p d\pi^*(x, y) + \int_A \|y - \mathcal{H}G_\Theta(y)\|^q d\pi^*(x, y)$$

Furthermore, we use disintegration theorem (Simmons, 2012) to split the joint measure π^* as follows:

$$d\pi^*(x, y) = \pi^*(dy|x) \mu(dx) = \pi^*(dx|y) \nu(dy),$$

where $\pi^*(\mathcal{Y}|x) = 1$ μ-a.e. and $\pi^*(\mathcal{X}|y) = 1$ ν-a.e. Then, we have

$$
\begin{aligned}
\int_B \|x - G_\Theta(\mathcal{H}x)\|^p d\pi^*(x, y) &= \int_B \|x - G_\Theta(\mathcal{H}x)\|^p \pi^*(dy|x) \mu(dx) \\
&= \int_{\mathcal{X}} \int_{y = \mathcal{H}x} \|x - G_\Theta(\mathcal{H}x)\|^p \pi^*(dy|x) \mu(dx) \\
&= \int_{\mathcal{X}} \|x - G_\Theta(\mathcal{H}x)\|^p \int \mathbf{1}_{y = \mathcal{H}x} \pi^*(dy|x) \mu(dx) \\
&= \int_{\mathcal{X}} \|x - G_\Theta(\mathcal{H}x)\|^p \rho(x) \mu(dx)
\end{aligned}
$$

where $\mathbf{1}_S$ denotes the indicator function for the set S and $\rho(x) = \int \mathbf{1}_{y = \mathcal{H}x} \pi^*(dy|x)$. In a similar fashion, with $\sigma(y) = \int \mathbf{1}_{x = G_\Theta(y)} \pi^*(dx|y)$, we have

$$\int_A c(x, y; \Theta, \mathcal{H}) d\pi^*(x, y) = \int_{\mathcal{Y}} \|y - \mathcal{H}G_\Theta(y)\|^q \sigma(y) \nu(dy) .$$

This concludes the proof. $\square$
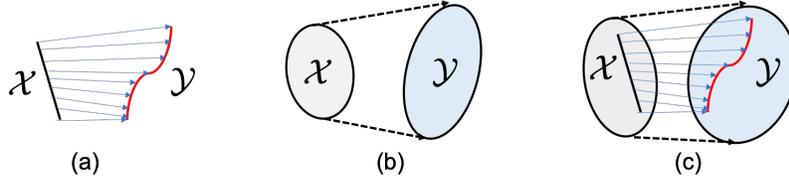
Figure 1: (a) Supervised learning with matched references, (b) optimal transport with absolute continuous measures, and (c) optimal transport having measures with singularities. CycleGAN corresponds to the scenario (c).

Here, care should be taken regarding the measurable functions ρ and σ. For the case of supervised learning, there exists perfectly matched reference as shown in Fig. 1(a) so that $\rho(x) = 1$; μ-a.e. and $\sigma(y) = 1$; ν-a.e. In this case, the $\ell_{OT'} = 0$. On the other hand, if the joint distribution $\pi(x, y)$ is absolutely continuous as shown in Fig. 1(b), then $\rho(x) = 0$; μ-a.e. and $\sigma(y) = 0$; ν-a.e. In this case, $\ell_{cycle} = 0$. Most interesting case arises when the joint measure $\pi(x, y)$ has singularities on the set of $A \cup B = \{y = \mathcal{H}x \text{ or } x = G_\Theta(y)\}$ (see Fig. 1(c)). In fact, the unsupervised learning cares about such situation where once the mappings are found, significant portion of the data in $\mathcal{X}$ and $\mathcal{Y}$ can be paired with the mappings, even though there are still remaining unpaired data. This is when the cycle consistency terms plays the important role in unsupervised learning.

The remaining terms in (10), which corresponds to the GAN term, can be obtained as follows:

Proposition 2. *For $p = q = 1$, the cost $\ell_{OT'}(\Theta, \mathcal{H})$ in (12) can be equivalently represented as*

$$\ell_{OT'}(\Theta, \mathcal{H}) = \max_\varphi \int_\mathcal{X} \varphi(x)d\mu(x) - \int_\mathcal{Y} \varphi(G_\Theta(y))d\nu(y) \tag{14}$$

$$+ \max_\psi \int_\mathcal{Y} \psi(y)d\nu(y) - \int_\mathcal{X} \psi(\mathcal{H}x)d\nu(x) \tag{15}$$

where φ and ψ are 1-Lipschitz Kantorovich potential functions.

Proof. This is a just simple corollary of the original Kantorovich's duality formulation proof and the classical results of optimal transport on the restricted measure. We start with the first term of $l_{OT}(\Theta, \mathcal{H})$ in (12). Specifically, we have

$$\int_{\mathcal{X} \times \mathcal{Y} \setminus (A \cup B)} \|G_\Theta(y) - x\|^p \pi^*(dx, dy) = \max_\varphi \int_\mathcal{X} \varphi(x)\mu'(dx) + \int_\mathcal{Y} \varphi^c(G_\Theta(y))\nu'(dy)$$

where μ' and ν' are marginals of the restriction of optimal transportation plan on the restricted set $\mathcal{X} \times \mathcal{Y} \setminus (A \cup B)$. Now, using Proposition 3 in Appendix, we have

$$\int_{\mathcal{X} \times \mathcal{Y} \setminus (A \cup B)} \|G_\Theta(y) - x\|\pi^*(dx, dy) = \max_\varphi \int_\mathcal{X} \varphi(x)\mu(dx) + \int_\mathcal{Y} \varphi^c(G_\Theta(y))\nu(dy)$$

$$= \max_{\varphi \in L^1(\mu)} \int_\mathcal{X} \varphi(x)\mu(dx) - \int_\mathcal{Y} \varphi(G_\Theta(y))\nu(dy)$$

for the non-restricted marginals μ and ν, where for the last equality we use $\varphi^c = -\varphi$ for $p = 1$ when φ is 1-Lipschitz (Villani, 2008). Using the same technique, the second term in (12) can be equivalently represented as

$$\int_{\mathcal{X} \times \mathcal{Y} \setminus (A \cup B)} \|y - \mathcal{H}x\|d\pi(x, y) = \max_{\psi \in L^1(\nu)} \int_\mathcal{Y} \psi(y)d\nu(y) - \int_\mathcal{X} \psi(\mathcal{H}x)d\nu(x).$$

By collecting terms, we conclude the proof. □

Now, the final step is implementing the Kantorovich potential using CNNs with parameters Φ and Ξ, i.e. $\varphi := \varphi_\Phi$ and $\psi := \psi_\Xi$. By collecting all terms in the transportation cost $\mathbb{T}(\Theta, \mathcal{H})$ together, the dual formulation results in a cycleGAN cost function:

$$\min_{\Theta, \mathcal{H}} \mathbb{T}(\Theta, \mathcal{H}) = \min_{\Theta, h} \max_{\Phi, \Xi} \ell(\Theta, \mathcal{H}; \Phi, \Xi) \tag{16}$$

where

$$\ell(\Theta, \mathcal{H}; \Phi, \Xi) = \ell_{cycle}(\Theta, \mathcal{H}) + \ell_{GAN}(\Theta, \mathcal{H}; \Phi, \Xi) \tag{17}$$

where $\ell_{cycle}(\Theta, \mathcal{H})$ denotes the cycle consistent loss in (13) and $\ell_{GAN}(\Theta, h; \Phi, \Xi)$ is the GAN loss given by:

$$\ell_{GAN}(\Theta, \mathcal{H}; \Phi, \Xi) \tag{18}$$
$$= \int_{\mathcal{X}} \varphi_\Phi(x) d\mu(x) - \int_{\mathcal{Y}} \varphi_\Phi(G_\Theta(y)) d\nu(y) + \int_{\mathcal{Y}} \psi_\Xi(y) d\nu(y) - \int_{\mathcal{X}} \psi_\Xi(\mathcal{H}x) d\nu(x)$$

Here, the Kantorovich 1-Lipschitz potential $\varphi := \varphi_\Phi$ and $\psi := \psi_\Xi$ corresponds to the Wasserstein GAN discriminators. Specifically, φ_Φ tries to find the difference between the true image x and the generated image $G_\Theta(y)$, whereas $\psi := \psi_\Xi$ attempts to find the fake measurement data that are generated by the synthetic measurement procedure $\mathcal{H}x$. Furthermore, if ρ and σ are constant almost everywhere, then they become simple scaling factors for the cycle-consistency terms:

$$\ell_{cycle}(\Theta, \mathcal{H}) = \rho \int_{\mathcal{X}} \|x - G_\Theta(\mathcal{H}x)\|^p d\mu(x) + \sigma \int_{\mathcal{Y}} \|y - \mathcal{H}G_\Theta(y)\|^q d\nu(y) \tag{19}$$

Note that our cost formulation using (17) with (18) and (19) is more general compared to the standard cycleGAN, since a general form of measurement data generator $\mathcal{H}x$ can be used. In the following section, we will show how to choose $\mathcal{H}$ depending on specific applications.

4 APPLICATIONS TO INVERSE PROBLEMS

4.1 ACCELERATED MRI

In accelerated MRI, the goal is to recover high quality MR images from sparsely sampled k-space data to reduce the acquisition time. This problem has been extensively studied using compressed sensing (Lustig et al., 2007), but recently deep learning approaches have been the main research interest due to its excellent performance at significantly reduced run time complexity (Hammernik et al., 2018; Han & Ye, 2019). A standard method for neural network training for accelerated MRI is based on the supervised learning, where the MR images from fully sampled k-space data is used as reference and subsampled k-space data is used as the input for the neural network. Unfortunately, in accelerated magnetic resonance imaging (MRI), high-resolution fully sampled k-space data is very difficult to acquire due to the long scan time. Therefore, the need for unsupervised learning without matched reference data is increasing.

In accelerated MRI, the forward measurement model can be described as

$$\hat{x} = \mathcal{P}_\Omega \mathcal{F} x + w \tag{20}$$

where $\mathcal{F}$ is the 2-D Fourier transform and $\mathcal{P}_\Omega$ is the projection to Ω that denotes k-space sampling indices. To implement every step of the algorithm as the image domain processing, (20) can be converted to the image domain forward model by applying the inverse Fourier transform:

$$y = \mathcal{F}^{-1} \mathcal{P}_\Omega \mathcal{F} x + \mathcal{F}^{-1} w \tag{21}$$

This results in the following cost function for the PLS formulation:

$$c(x, y; \Theta) = \|y - \mathcal{F}^{-1} \mathcal{P}_\Omega \mathcal{F} x\| + \|G_\Theta(y) - x\| \tag{22}$$

Usually, the sampling mask Ω is known so that the forward mapping for the inverse problem is deterministic.

The schematic diagram of the associated cycleGAN architecture is illustrated in Fig. 2, whose generator and discriminator architecture are shown in Fig. 6 in Appendix. Note that we just need as single generator, as the mapping from the clean to aliased images is deterministic for a given sampling pattern. As for the loss function, we use (17) with the $\sigma = \rho = 0.5$ for the cycle consistency term in (19). For GAN implementation, we use the Wasserstein GAN in (18) with Lipschitz penalty loss (Gulrajani et al., 2017) to enforce that the Kantorovich potential become 1-Lipschitz. The detailed description of the training procedure is provided in Appendix. The reconstruction results in Fig. 3 clearly showed that the proposed unsupervised learning method using dedicated cycleGAN successfully recovered fine details without matched references.
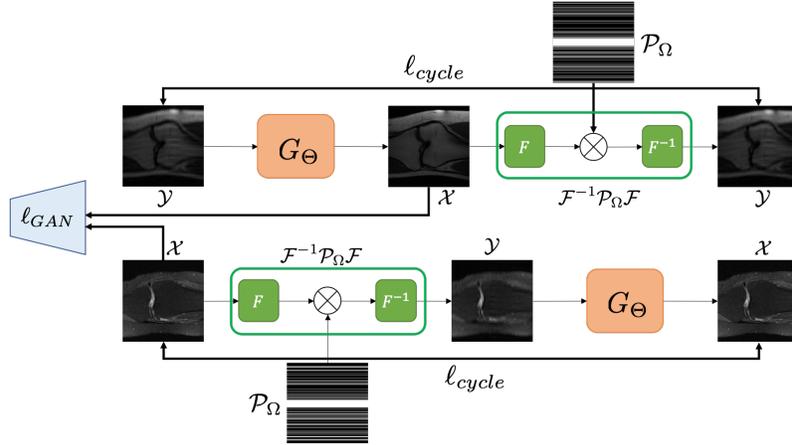
Figure 2: Proposed cycleGAN architecture for accelerated MRI with 1-D downsampling patterns.

4.2 DECONVOLUTION MICROSCOPY

Deconvolution microscopy is extensively used to improve the resolution of the widefield fluorescent microscopy. Recently, CNN approaches have been extensively studied as fast and high performance alternatives. Unfortunately, the CNN approaches usually require matched high resolution images for supervised training. Mathematically, a blurred measurement can be described as

$$y \quad = \quad h * x + w \,, \tag{23}$$

where h is the point spread function (PSF). If the PSF h is not known, which is often referred to as the blind deconvolution problem, both the unknown PSF h and the image x should be estimated. This results in the following cost function for the PLS formulation:

$$c(x, y; \Theta, h) = \|y - h * x\| + \|G_\Theta(y) - x\| \tag{24}$$

The corresponding cycleGAN architecture is illustrated in Fig. 4. In contrast to the conventional cycleGAN approaches that require two generators, the proposed cycleGAN approach needs only a single generator and the blur image generation can be done using a linear convolution layer corresponding to PSF, which significantly improves the robustness of network training. The detailed description of the network architecture and training are given in Appendix.
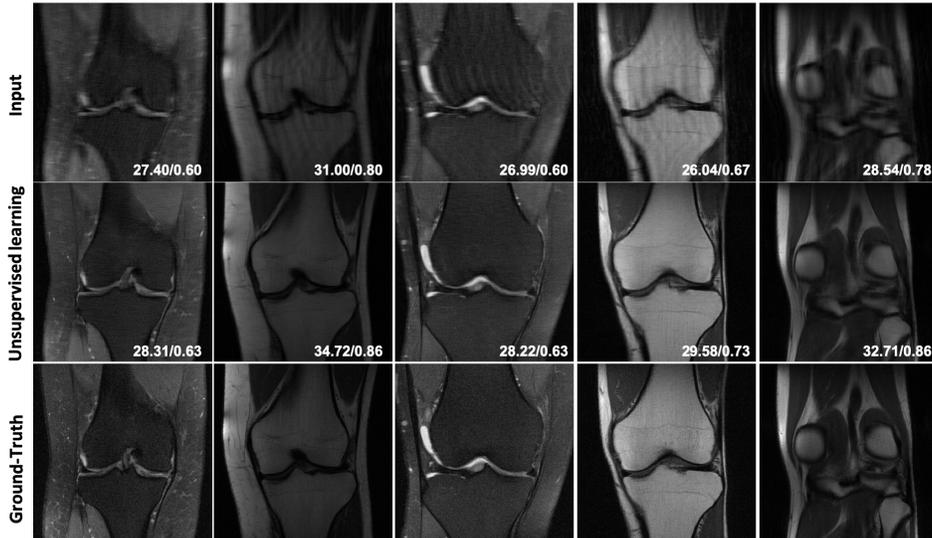
Figure 3: Unsupervised learning results for accelerated MRI using proposed cycleGAN. The values in the corner are PSNR/SSIM values.

7
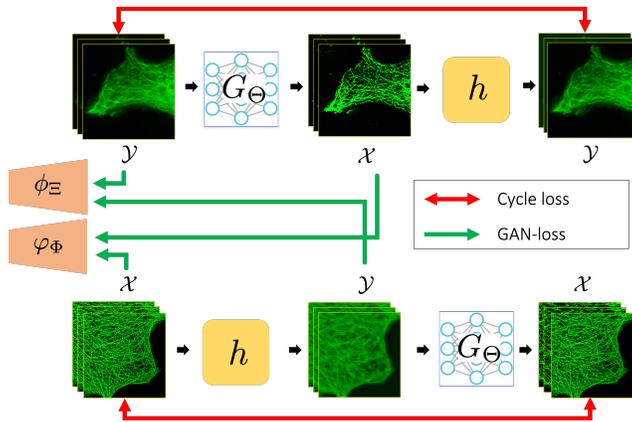
Figure 4: Proposed cycleGAN architecture with a blur kernel for deconvolution microscopy. Here, G_Θ denotes the CNN-based low-resolution to high-resolution generator. The blur generator is composed of a linear blur kernel h.

Fig. 5 show lateral views of deconvolution results of microtubule samples by various methods. Here, input images are degraded by blur and noises. The supervised learning and the standard cycleGAN with two generators showed better contrast and removed blur; however, the structural continuity was not preserved. On the other hand, in our cycleGAN approach, blurs and noise were successfully removed and preserved the continuity of the structure.
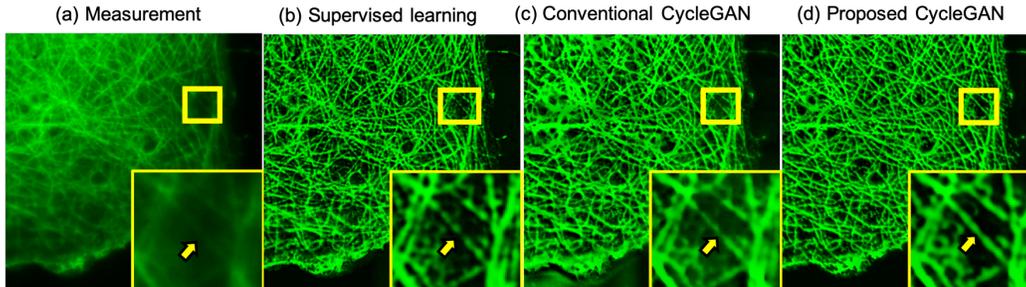
Figure 5: Comparison of reconstruction results by various methods: (a) Blurred image measurements, (b) supervised learning, (c) the conventional cycleGAN, and (d) the proposed cycleGAN. The ROIs (marked yellow) show the area for the enlarged parts.

5 CONCLUSIONS

In this paper, we presented a general design principle for cycleGAN architecture for various inverse problems. Specifically, the proposed architecture was derived as a dual formulation of an optimal transport problem that uses the penalized least squares cost as the transport cost between two measures supported on the measurement data and the images. As proofs of concept, we designed novel cycleGAN architecture for accelerated MRI and deconvolution microscopy examples, providing accurate reconstruction results without any matched reference data. Given the generality of our design principle, we believe that our method can be an important platform for unsupervised learning for inverse problems.

REFERENCES

Hemant K Aggarwal, Merry P Mani, and Mathews Jacob. MoDL: Model-based deep learning architecture for inverse problems. *IEEE transactions on medical imaging*, 38(2):394–405, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

Subhasis Chaudhuri, Rajbabu Velmurugan, and Renu Rameshan. Blind deconvolution methods: A review. In *Blind Image Deconvolution*, pp. 37–60. Springer, 2014.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pp. 424–432. Springer, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *Advances in neural information processing systems*, pp. 5767–5777, 2017.

Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated MRI data. *Magnetic resonance in medicine*, 79(6):3055–3071, 2018.

Yoseob Han and Jong Chul Ye. k-Space Deep Learning for Accelerated MRI. *IEEE Transactions on Medical Imaging (in press), also available as arXiv preprint arXiv:1805.03779*, 2019.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Eunhee Kang, Junhong Min, and Jong Chul Ye. A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction. *Medical physics*, 44(10), 2017.

Eunhee Kang, Hyun Jung Koo, Dong Hyun Yang, Joon Bum Seo, and Jong Chul Ye. Cycle-consistent adversarial denoising network for multiphase coronary CT angiography. *Medical physics*, 46(2):550–562, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. Conditional cyclegan for attribute guided face image generation. *arXiv preprint arXiv:1705.09966*, 2017.

Michael Lustig, David Donoho, and John M Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magn. Reson. Med.*, 58(6):1182–1195, 2007.

James G McNally, Tatiana Karpova, John Cooper, and José Angel Conchello. Three-dimensional imaging by deconvolution microscopy. *Methods*, 19(3):373–385, 1999.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

Pinaki Sarder and Arye Nehorai. Deconvolution methods for 3-D fluorescence microscopy images. *IEEE Signal Processing Magazine*, 23(3):32–45, 2006.

David Simmons. Conditional measures and conditional expectation; rohlin's disintegration theorem. *Discrete & Continuous Dynamical Systems-A*, 32(7):2565–2582, 2012.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Matthew J Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J Geras, Joe Katsnelson, Hersh Chandarana, et al. fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv preprint arXiv:1811.08839*, 2018.

Jiawei Zhang, Jinshan Pan, Wei-Sheng Lai, Rynson WH Lau, and Ming-Hsuan Yang. Learning fully convolutional networks for iterative non-blind deconvolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3817–3825, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

A APPENDIX

A.1 ADDITIONAL THEOREMS AND LEMMAS

Theorem 2 (Optimality is inherited by restriction)**.** *(Villani, 2008, Theorem 4.6, p.46) Under the same assumption of Kantorovich duality theorem, let π^* be an optimal transport plan, and $\tilde{\pi}$ be a nonnegative measure on $\mathcal{X} \times \mathcal{Y}$, such that $\tilde{\pi} \leq \pi^*$ and $\tilde{\pi}[\mathcal{X} \times \mathcal{Y}] > 0$. Let $\pi' := \frac{\tilde{\pi}}{\tilde{\pi}[\mathcal{X} \times \mathcal{Y}]}$, μ', ν' be the marginals of π'. Then optimality is inherited to its restriction, i.e., π' is an optimal transference plan between μ' and ν'.*

Theorem 3 (Restriction for the Kantorovich duality)**.** *(Villani, 2008, Theorem 5.19, p.75-p.76) Under the same assumption of Kantorovich duality theorem, let π^* be an optimal transport plan and φ^* be an optimal potential that attain the maximum, and $\tilde{\pi}$ be a measure on $\mathcal{X} \times \mathcal{Y}$, such that $0 \leq \tilde{\pi} \leq \pi^*$ and $Z := \tilde{\pi}[\mathcal{X} \times \mathcal{Y}] > 0$. Let $\pi' = \tilde{\pi}/Z$, and μ', ν' be the marginals of π'. Then there exists φ' such that $\varphi' = \varphi^*$ on $proj_{\mathcal{X}}(supp(\pi^*))$, and φ' solves the dual Kantorovich problem between $(\mathcal{X}, \mu')$ and $(\mathcal{Y}, \nu')$.*

We are now ready to provide our duality result for the restricted measure.

Proposition 3. *Let*

$$\ell_{OT'} := \min \pi \int_{\mathcal{X} \times \mathcal{Y} \setminus (A \cup B)} c(x,y) d\pi(x,y)$$

$$\ell_{OT} := \min \pi \int_{\mathcal{X} \times \mathcal{Y}} c(x,y) d\pi(x,y)$$

Then, we can replace $\ell_{OT'}$ with ℓ_{OT} in the sense that both of them share the same optimal Kantorovich potential. More specifically, if φ^* is the optimal Kantorovich potential, which satisfies

$$\ell_{OT} = \int_{\mathcal{X}} \varphi^*(x) d\mu(x) + \int_{\mathcal{Y}} (\varphi^*)^c(y) d\nu(y),$$

then

$$\ell_{OT'} = \int_{\mathcal{X}} \varphi^*(x) d\mu'(x) + \int_{\mathcal{Y}} (\varphi^*)^c(y) d\nu'(y),$$

where μ' and ν' are marginals of the restriction of optimal transference plan.

Proof. Let $\tilde{\pi} = \pi^*|_{\mathcal{X} \times \mathcal{Y} \setminus (A \cup B)}$. Then, $\tilde{\pi}$ is nonnegative measure on $\mathcal{X} \times \mathcal{Y}$. When $\tilde{\pi}(\mathcal{X} \times \mathcal{Y}) = 0$, $\ell_{OT'} = 0$, and $\mu' = 0, \nu' = 0$, so the statement trivially holds. Suppose $\tilde{\pi}(\mathcal{X} \times \mathcal{Y}) > 0$. Then by lemma, there exists φ' such that $\varphi' = \varphi^*$ on $proj_{\mathcal{X}}(supp(\pi^*))$, and φ' solves the dual Kantorovich problem between $(\mathcal{X}, \mu')$ and $(\mathcal{Y}, \nu')$. □

A.2 EXPERIMENTAL DETAILS

A.2.1 ACCELERATED MRI

We use dataset for fastMRI challenge (Zbontar et al., 2018) for our experiments. This dataset is composed of MR images of knees. We extracted 3500 MR images from fastMRI single coil validation

set. Then, 3000 slices are used for training/evaluation, and 500 slices are used for test. These MR images are fully sampled images, so we make undersampled images by a randomly subsampling k-space lines. The acceleration factor is four, and autocalibration signal (ACS) region contains 4% of k-space lines. Each slice is normalized by standard deviation of the magnitude of each slice. To handle complex values of data, we concatenate real and imaginary values along the channel dimension. Each slice has different size, so we use only single batch. The images are center-cropped to the size of 320×320, and then the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) values are calculated.

We use U-Net generator to reconstruct fully sampled MR images from undersampled MR images as shown in Fig. 6(a). Our generator consists of 3×3 convolution, Instance normalization, and leaky ReLU operation. Also, there are skip-connection and pooling layers. At the last convolution layer, we do not use any operation. Our discriminator is same as the discriminator of original CycleGAN. We use PatchGAN (Isola et al., 2017), so the discriminator classifies inputs at patch scales. The discriminator also consists of convolution layer, instance normalization, and leaky ReLU operation as shown in Fig. 6(b). We use Adam optimizer to train our network, with momentum parameters $\beta_1 = 0.5, \beta_2 = 0.9$, and learning rate of 0.0001. The discriminator is updated 5 times for every generator updates. We use batch size of 1, and trained our network during 100 epochs. Our code was implemented by TensorFlow.
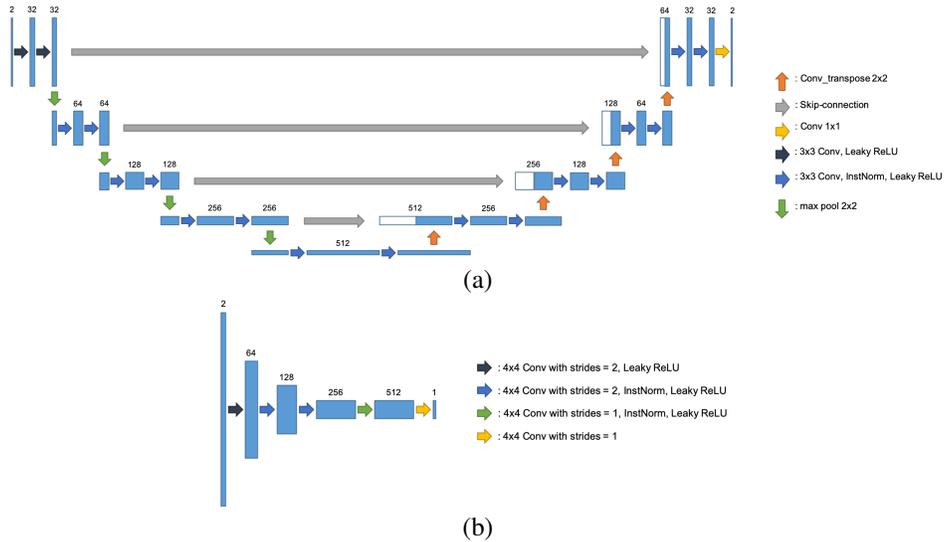
Figure 6: Proposed network architecture for (a) generator and (b) discriminator.
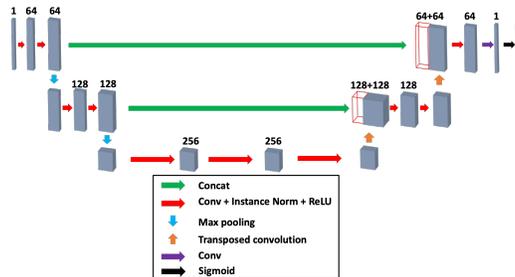
Figure 7: A modified 3D U-net architecture for our high-resolution image generator.

A.2.2 DECONVOLUTION MICROSCOPY

The network architecture of the high resolution image generator G_Θ from the low-resolution image is a modified 3D-Unet (Çiçek et al., 2016) as shown in Fig 7. Our U-net structure consists of
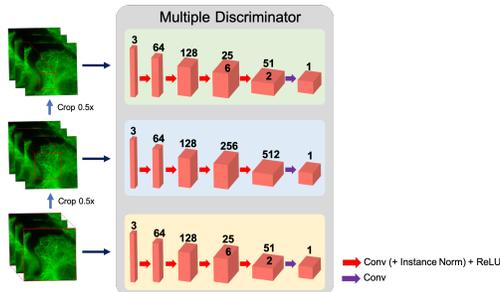
Figure 8: Multi-PatchGANs discriminator architecture.

contracting and expanding paths. The contracting path consists of the repetition of the following blocks: 3D conv- Instance Normalization (Ulyanov et al., 2016)- ReLU. Here, the generator has symmetric configuration so that both encoder and decoder have the same number of layers, i.e. $\kappa = 7$. Throughout the network, the convolutional kernel dimension is $3 \times 3 \times 3$. There exists a pooling layer and skipped connection for every other convolution operations. To enhance the image contrast, we add additional sigmoid layer at the end of U-Net. On the other hand, the low-resolution image generator from high resolution input is based on a single 3D convolution layer that models a 3D blurring kernel. The size of the 3D PSF modeling layer is chosen depending on the problem set by considering their approximate PSF sizes. In this paper, the size of the 3D PSF layer is set to $20 \times 20 \times 20$.

As for the discriminators, we follow the original CycleGAN that uses multi-PatchGANs (mPGANs) (Isola et al., 2017), where each discriminator has input patches with different sizes used. As shown in Fig 8, it consist of three independent discriminators. Each discriminator takes patches at different sizes: original, and half, and quarter size patches.

A total 18 epi-fluorescent (EPF) microscopy images of tubulin with a size of $512 \times 512 \times 30$ were used for training, and one for validation. As for unmatched sharp image volume, we used deblurred image generated by utilizing a commercial software AutoQuant X3 (Media Cybernetics, Rockville). The EPF volume depth was increased to 64 using the reflection boundary condition. Due to GPU memory limitations, the EPF volume was split into $64 \times 64 \times 64$ size patches. For data augmentation, rotation, flip, translation, and scale were imposed on the input patches. We normalized the patches and set them to [0,1]. Adam optimizer (Kingma & Ba, 2014) was also used for training. The learning rate was initially set to 0.0001, which is decreased linearly after 40 epoch, and the total number of epoch was 200 epoch. For the optimizer, we used only a single batch.