Optimal Domain Translation

Emmanuel de Bézenac [*1] **Ibrahim Ayed** [*12] **Patrick Gallinari** [13]

Abstract

Domain Translation consists in transforming elements from one domain to another, in such a way that their semantics are conserved across domains. This task is very general, and has abundant applications across diverse fields such as style transfer, machine translation, transfer learning, inverse problems, music synthesis, etc.. However, the lack of paired data from each domain has led the community to consider solving the previous problem with unpaired data only. Despite impressive successes for the unpaired problem, there remains little theoretical understanding. After showing that the common approach for the unpaired problem is ill-posed, we reformulate the problem as a transport problem between two probability distributions, explicitly enforcing semantics to be conserved across domains, which makes it equivalent to the problem of optimal transport, thus yielding both existence and unicity of the transformation. We then consider a dynamical formulation for the transportation problem which allows us to build our model as an ODE flow of minimal energy. Finally, we evaluate our method on toy examples as well as classical datasets.

Given an input and output domain, *domain translation* consists in learning a mapping between the elements of the two domains, so as to retrieve *semantically meaningful* pairings between the two domains. For example, if the domains were natural photographs and art paintings, such a pairing could refer to couples of natural photographs and paintings that describe the same underlying scene. A wide range of problems could be formulated as translation, including image-to-image (Isola et al., 2016) or video-to-video (Wang et al., 2018) translation, image captioning (Zhang et al., 2016), natural language translation (Bahdanau et al., 2015), music synthesis (Mor et al., 2018), and many more. The problem has been initially tackled in the supervised (or paired) setting, where we have access to data with labelled pairings. Considering the unsupervised (or unpaired) setting, where no paired data is available, allows to tackle an even wider range of problems, and has recently motivated several works from the community with applications in pose estimation (Kanazawa et al., 2018), speech-to-text (Chung et al., 2018), data augmentation (Lee et al., 2018) or text style transfer (Subramanian et al., 2018). One of the seminal work in this direction has been the CycleGAN model proposed by (Zhu et al., 2017) for image-to-image translation.

If one is to calculate a correspondance between two domains, *inversibility* is an important feature of any candidate solution. However, the problem is still an ill-posed one, with many possible solutions as there are many inversible mappings, possibly an infinity, which can associate two distributions. This means that an additional criterion has to be taken into account, one which would ideally select a unique solution for UDT.

Our first contribution is to reformulate UDT as an optimal transport problem (OT). This will guarantee the existence and unicity of a solution. In order to derive an efficient algorithm, we then use a dynamical formulation of OT in order to introduce an algorithm for solving the transport problem. This new formulation leads to learning trajectories with minimal length (geodesics) in the measure space where the source and target distributions lie. We parameterize these trajectories as ODE flows controlled by a constant velocity fields which naturally yields the desired transformation and its inverse. Finally, we evaluate our method over different datasets and compare it to previous baselines. Our main contributions are then following:

- We cast the unsupervised domain translation problem in an Optimal Transport setting thus providing a new way to attack UDT.
- We make use of a dynamical formulation of OT in order to frame the learning problem to build a continuous model and describe the corresponding algorithm.
- We empirically investigate the performance of our model as well as its different properties and features as compared to other models.

[*]Equal contribution [1]Sorbonne Université, UMR 7606, LIP6, F-75005 Paris, France [2]Theresis lab, Thales, Thales Research & Technology Route Départementale, 91120 Palaiseau [3]Criteo AI Lab, Paris, France. Correspondence to: Ibrahim Ayed <ibrahim.ayed@lip6.fr>, Emmanuel de Bézenac <emmanuel.de-bezenac@lip6.fr>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

1. Unpaired Domain Translation as Optimal Transport

1.1. Definition and properties

We start by stating our problem:

Definition 1.1. Unsupervised Domain Translation (UDT) Given two domains represented by probability distributions α and β, the problem of domain translation consists in finding couplings between α and β, *i.e.* mappings $T\colon \mathcal{A} \to \mathcal{B}$ and $S\colon \mathcal{B} \to \mathcal{A}$, in such a way that these couplings are *semantically meaningful*.

A common approach for UDT is based on (Zhu et al., 2017). Therefore, many approaches (Lample et al., 2018; Yuan et al., 2018; Chung et al., 2018; Choi et al., 2018) attempt to enforce a *coherence* constraint, acting upon the the output distributions induced by the mappings T and S, and an *inversibility* constraint, applied on the transformed elements itself:

– ***Coherence***[1]: $T_\sharp \alpha = \beta$, and $S_\sharp \beta = \alpha$.

– ***Inversibility***[2]: $S \circ T \overset{\alpha-a.s.}{=} \mathrm{id}$, and $T \circ S \overset{\beta-a.s.}{=} \mathrm{id}$,

For instance, the *CycleGan* model (Zhu et al., 2017) minimizes a weighted combination of a coherence loss $D(T_\sharp \alpha, \beta) + D(S_\sharp \beta, \alpha)$, where D corresponds to an adversarial loss, and a cycle-consistent loss $\|S \circ T - \mathrm{id}\|_{L^1(\alpha)} + \|T \circ S - \mathrm{id}|_{L^1(\beta)}$, using the norm of L^1.

However, it is important to stress the fact that those two constraints are not sufficient to ensure a unique solution to UDT. This can be easily seen in the discrete case, when the two distributions are simply sums of Dirac distributions: any arbitrary pairing between both domains satisfies both the *inversibility* and *coherence* constraints. This can be easily extended to the continuous setting when the two distributions are absolutely continuous and admit a density: We can prove that there exists infinitely many mappings satisfying both constraints.

1.2. Optimal Transport

In the following, we consider that we have access to a finite number of samples of both distributions distributions α, β which are assumed to be absolutely continuous *w.r.t.* Lebesgue measure.

Let us consider an abstract cost function $c(x, y)$, low when x and y have the same semantics, and high otherwise. For example, in the case where x, y are semantic representations of domain elements, c could correspond to the euclidean distance between x and y: $c(x, y) = \|x - y|_2^2$. The mapping of x through $f\colon \mathcal{R}^d \to \mathcal{R}^d$ then incurs a cost $c(x, f(x))$. Hence, a *semantic preserving mapping* is a mapping with a low cost over the elements on the support of distribution μ, *i.e.* a low value for the integral $\int_{\mathcal{R}^d} c(x, f(x))\, \mathrm{d}\mu(x)$.

Out of all mappings that satisfy the *invertibility* and *coherence* conditions, we are now looking for mappings T and S that are of minimal cost:

$$\begin{aligned} \underset{T,S}{\text{minimize}} \quad & \int_{\mathcal{R}^d} c(x, T(x))\, \mathrm{d}\alpha(x) \;+\; \int_{\mathcal{R}^d} c(y, S(y))\, \mathrm{d}\beta(y) \\ \text{subject to} \quad & T_\sharp \alpha = \beta, \quad S_\sharp \beta = \alpha, \\ & S \circ T \overset{\alpha-a.s.}{=} \mathrm{id}, \quad T \circ S \overset{\beta-a.s.}{=} \mathrm{id} \end{aligned} \tag{1}$$

Consider now the classical *Monge formulation* of OT (MP):

$$\begin{aligned} \underset{T}{\text{minimize}} \quad & \mathcal{C}(T) = \int_{\mathcal{R}^d} c(x, T(x))\, \mathrm{d}\alpha(x) \\ \text{subject to} \quad & T_\sharp \alpha = \beta \end{aligned} \tag{2}$$

This is the modern formulation of the original OT problem initially stated by Gaspard Monge, and is the founding stone of the celebrated Monge-Kantorovitch theory. The following important result not only states that solving equation 1 actually boils down to solving equation 2, but it also provides us with existence and unicity of the optimal mappings T and S:

Theorem 1. *Let α, β absolute continuous measures. If $c(x, y) = h(x - y)$ where h is strictly convex, then there* ***exists a unique*** *couple (T, S) of transformations such that:*

- *$T_\sharp \alpha = \beta$, and $S_\sharp \beta = \alpha$,*
- *$\mathcal{C}(T)$ is minimal, and S is the minimal transport from β to α.*
- *$T \circ S \overset{\beta-a.s.}{=} \mathrm{id}$, and $S \circ T \overset{\alpha-a.s.}{=} \mathrm{id}$,*

Its proof can be found in (Santambrogio, 2015), among other classical OT references.

This result has several implications:

– All the constraints in equation 1 are naturally verified by the optimal maps T and S, without having to enforce them explicitly,

– existence and uniqueness of the optimal maps may actually shed light on the recent success of unpaired domain translation approaches.,

[1]The *push-forward* $f_\sharp \rho$ is defined as $f_\sharp \rho(B) = \rho(f^{-1}(B))$, for any measurable set B.

[2]Notation $f \overset{\mu-a.s.}{=} g$ expresses that $\int_B f \,\mathrm{d}\mu = \int_B g \,\mathrm{d}\mu$, for any measurable set B.

– solving on both T and S is not necessary, as solving only on T yields the same result,

– we can consider equation 2 instead of equation 1.

However, this theorem is not constructive in the sense that it does not give us any clue about how we could retrieve such optimal solutions. In the following, we use a dynamical framing for the problem which yields naturally a model for solving it.

2. Neural dynamical formulation of Optimal Transport

2.1. Dynamic Formulation for Optimal Transport

Instead of directly pushing α to β, it is possible to view α and β as points in a space of measures, and consider trajectories from α to β. Thus, a way to transport the probability mass from α to β is a curve between two points on this space. The curve corresponding to the optimal mapping is then the *shortest* one, in other words it is the *geodesic curve* between the two points.

More formally, let us introduce the *Wasserstein* metric space $\mathbb{W}_p(\mathcal{R}^d)$, *i.e.* the space of measures of $\mathcal{R}^d$ with finite p-th moment endowed with the Wassertein distance:

$$W_p(\mu, \nu) = \min_{T_\sharp \mu = \nu} \mathcal{C}(T)^{\frac{1}{p}}$$

when costs of the form $c(x, y) = \|x - y\|^p$ are considered, for some integer $p \geq 2$. As $\mathbb{W}_p(\mathcal{R}^d)$ is a space of measures, α and β are seen as points of this space of measures, and thus, any continuous path linking both distributions defines a gradual transformation from α to β.

The following result (from Theorem 5.27 of (Santambrogio, 2015)) motivates the dynamic formulation of OT:

Proposition 1. *$\mathbb{W}_p$ is a geodesic space, meaning that, for any measures $\mu, \nu \in \mathbb{W}_p$ there exists a geodesic curve $(\mu_t)_{t\in[0,1]}$ between μ and ν.*

The following theorem, for which a proof can be found in (Santambrogio, 2015), the famed *Benamou-Brenier formula* makes the analogy between the motion of fluid mass from one configuration to another to the transport of probability masses, linking the geodesic curves to a minimal energy flow of a differential equation, gradually displacing probability mass from one domain to another.

Theorem 2. *Given α and β admitting densities w.r.t. the Lebesgue measure and $(\mu_t)_{t\in[0,1]}$ the geodesic curve with $\mu_0 = \alpha$ and $\mu_1 = \beta$, we can associate a vector field $v_t \in L^p(\mu_t)$ which solves the continuity equation[3]:*

$$\partial_t \mu_t + \nabla \cdot (\mu_t v_t) = 0$$

with:

$$W_p^p(\alpha, \beta) = \int_0^1 \|v_t\|^p_{L^p(\mu_t)} \mathrm{d}t$$

In other words, the geodesic curve $(\mu_t)_{t\in[0,1]}$ between both distributions, together with the minimal energy velocity vector field v solve the continuity equation. Moreover, its energy along this path is precisely equal to the Wasserstein distance $W_p^p(\alpha, \beta)$. If this vector field of minimal energy v could be obtained, probability mass could be displaced according to the flow defined by the continuity equation, and the geodesic curve could be retrieved. Thus, we can reformulate the problem as a problem of optimal control, where v is the control variable:

$$\begin{aligned} \underset{v}{\text{minimize}} \quad & \mathcal{C}^{\text{dyn}}(v) = \int_0^1 \|v_t\|^p_{L^p(\mu_t)} \mathrm{d}t \\ \text{subject to} \quad & \partial_t \mu_t + \nabla \cdot (\mu_t v_t) = 0, \mu_0 = \alpha, \mu_1 = \beta \end{aligned} \tag{3}$$

2.2. Solving the control problem

Having only access to samples from α and β, rather than using the continuity equation, we take the Lagrangian point of view which gives the equivalent equation:

$$\begin{cases} \partial_t \phi_t^x = v_t(\phi_t^x) \\ \phi_0^x = x \end{cases} \tag{4}$$

Moreover, as we intend to learn the control variable v, we parameterize it as v^θ which finally yields:

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \mathcal{C}^{\text{dyn}}(v^\theta) = \int_0^1 \|v_t^\theta\|^p_{L^p((\phi_t^\cdot)_\sharp \alpha)} \mathrm{d}t \\ \text{subject to} \quad & \partial_t \phi_t^x = v_t^\theta(\phi_t^x), \\ & \phi_0^\cdot = \mathrm{id}, \\ & (\phi_1^\cdot)_\sharp \alpha = \beta \end{aligned} \tag{5}$$

The constrained optimization problem in equation 5 is a continuous-time optimal control problem, similar to problems from fluid mechanics. This problem can be approached using gradient based optimization techniques. For instance, in our implementation using neural networks, we can either use back-propagation if a differentiable solver is taken to solve the forward equation or the adjoint method as detailed in (Chen et al., 2018).

Moreover, in practice, in order to enforce the constraint over the target, we use a discriminator.

3. Experiments

We evaluate the effectiveness of our methodology on:

- A simulated dataset of two gaussian distributions with different means and variances. In this dataset, we have

[3] ∂t is the partial derivative operator *w.r.t.* variable t, and $\nabla \cdot$ the divergence operator *w.r.t.* space.

Algorithm 1 Finding the Optimal Mapping

Input: Dataset of unpaired samples $x_{\mathcal{A}}$, $x_{\mathcal{B}}$, sampled from α, β

Guess initial parameters θ

while not converged **do**

  Randomly sample a mini-batch of $x_{\mathcal{A}}$, $x_{\mathcal{B}}$

  Solve $\partial_t \phi_x^t = v_\theta(\phi_x^t)$, $\phi_0^x = x_{\mathcal{A}}$, $t \in [0,1]$, for all $x_{\mathcal{A}}$

  Compute loss $\mathcal{C}$ on mini-batch

  Propagate gradient and compute estimate of $\frac{\mathrm{d}\mathcal{C}}{\mathrm{d}\theta}$

  Update θ in the steepest descent direction

end while

Output: Learned parameters θ.

sampled points from a normal distribution, and applied different linear transformations in order to obtain samples for each domain. As is shown in Table 1, we evaluate our capacity to retrieve the samples using our learned mappings.

- The classical CelebA dataset with the male and female distributions, for a high-dimensional example.

Table 1. Quantitative results for the different tested methods, on the two Gaussian distribution dataset. The first metric represents the transportation cost of the mapping. The second expresses how well the pairings between both domains are retrieved. Note that due to the fact that the coherence constraint is not necessarily verified, OT does not yield the smallest transportation cost.

METRIC	$\frac{1}{N}\sum_{x_{\mathcal{A}}} \lVert \phi_1^{x_{\mathcal{A}}} - x_{\mathcal{A}} \rVert_2^2$	$\frac{1}{N}\sum_{(x_{\mathcal{A}}, x_{\mathcal{B}})} \lVert \phi_1^{x_{\mathcal{A}}} - x_{\mathcal{B}} \rVert_2^2$
MLP	3.59	1.08
RESNET	4.03	0.005
FLOW	4.24	0.001
OT	4.12	0

3.1. Experiments details

For the first experiment, we parametrize v as a fully connected neural network with 2 hidden layers, ReLu (0.2) non-linearities and batch normalization. The discriminator is also a fully connected network with 3 hidden layers, ReLu non-linearities and spectral normalization. The dataset includes 1000 samples from each distribution.

For the second experiment, we choose a Convolutional Residual Network as a model for v, with 5 residual blocks and for the discriminator, the one from SAGAN (Zhang et al., 2018).

For both experiments, we used as solver explicit Euler steps

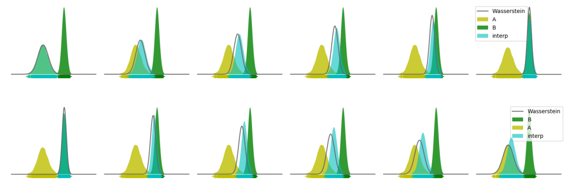

Figure 1. Gaussian distributions experiments: With our method, from left to right, we have $t = 0, 0.2, 0.4, 0.6, 0.8, 1$. Up we have the mapping from α to β, down the inverse.

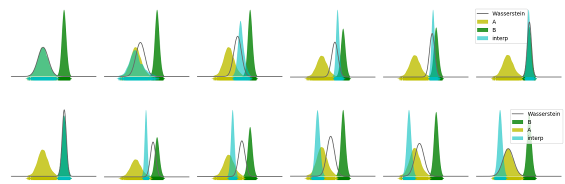

Figure 2. Gaussian distributions experiments: With a residual network, from left to right, we have $t = 0, 0.2, 0.4, 0.6, 0.8, 1$. Up we have the mapping from α to β, down the inverse.

with $\Delta t = \frac{1}{5}$. As this is a differentiable solver, we used back-propagation to calculate the gradient.

3.2. Results

Figure 1 shows that our method successfully learns a transport which is nearly optimal as well as its inverse. Let us stress the fact that, during train, we only had a loss over the mapping from α to β and nothing for the inverse, which is simply obtained by calculating the inverse flow. Figure 2 shows that a simple residual network, while giving a reasonable forward mapping, although the intermediate interpolations are less good, doesn't yield the right inverse.

Figures 3 and 4 show that our method can also be successfully applied to more complex high-dimensional datasets. Again, we successfully learn a convincing transformation as well as its inverse, as shown in figure 3, while figure 4 shows one of the advantages of our continuous method which yields smooth interpolations between the two distributions, as well as extrapolations.

4. Conclusion

In this short paper, we build an Optimal Transport framework for the ill-posed Unpaired Domain Translation problem. Using its dynamical formulation, we then build a model

Figure 3. CelebA experiment: Male to Female and back. From top to bottom, for male and female: input, reconstruction, after transport.

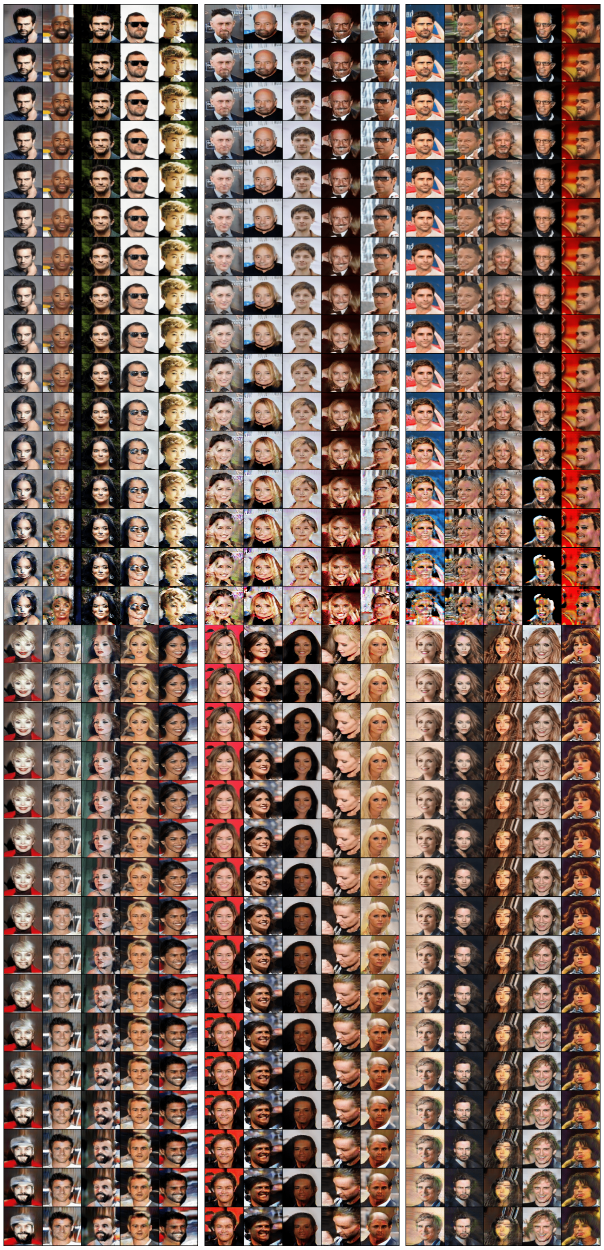

Figure 4. CelebA experiment: Interpolations and extrapolations of Male to Female, and back. The images correspond to images from $\{\mu_k\}_{k\in\{-\frac{5}{6},\ldots,\frac{10}{6}\}}$. Note that images from the input and output domain correspond to images 6 and 11, respectively. Same applies for female to male images, below.

which yields a provably unique solution and is tractable even in high-dimensional settings. Finally, we also show preliminary results with examples from low-dimensional, easy to visualize gaussian distributions as well as with classical image datasets from CelebA. The results correspond convincingly to the insights given by the theory. Moreover, it should be added that for models such as Resnets with residuals initialized to zero, which are often used for UDT, actually already yield a very small transportation cost and thus our theoretical considerations can also be seen as an explanation for the good results of those models.

References

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.0473.

Chen, T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *CoRR*, abs/1806.07366, 2018. URL http://arxiv.org/abs/1806.07366.

Choi, Y., Choi, M., Kim, M., Ha, J., Kim, S., and Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 8789–8797, 2018. doi: 10.1109/CVPR.2018.00916. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Choi_StarGAN_Unified_Generative_CVPR_2018_paper.html.

Chung, Y., Weng, W., Tong, S., and Glass, J. Unsupervised cross-modal alignment of speech and text embedding spaces. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pp. 7365–7375, 2018. URL http://papers.nips.cc/paper/7965-unsupervised-cross-modal-alignment-of-speech

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. *arxiv*, 2016.

Kanazawa, A., Black, M. J., Jacobs, D. W., and Malik, J. End-to-end recovery of human shape and pose. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 7122–7131, 2018. doi: 10.1109/CVPR.2018.00744. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Kanazawa_End-to-End_Recovery_of_CVPR_2018_paper.html.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=rkYTTf-AZ.

Lee, K., Kim, H., and Suh, C. Simulated+unsupervised learning with adaptive data generation and bidirectional mappings. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=SkHDoG-Cb.

Mor, N., Wolf, L., Polyak, A., and Taigman, Y. A universal music translation network. *CoRR*, abs/1805.07848, 2018. URL http://arxiv.org/abs/1805.07848.

Santambrogio, F. *Optimal transport for Applied Mathematicians: Calculus of Variations, PDEs and Modeling*. 2015.

Subramanian, S., Lample, G., Smith, E. M., Denoyer, L., Ranzato, M., and Boureau, Y. Multiple-attribute text style transfer. *CoRR*, abs/1811.00552, 2018. URL http://arxiv.org/abs/1811.00552.

Wang, T., Liu, M., Zhu, J., Liu, G., Tao, A., Kautz, J., and Catanzaro, B. Video-to-video synthesis. *CoRR*, abs/1808.06601, 2018. URL http://arxiv.org/abs/1808.06601.

Yuan, Y., Liu, S., Zhang, J., Zhang, Y., Dong, C., and Lin, L. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 701–710, 2018. doi: 10.1109/CVPRW.2018.00113. URL http://openaccess.thecvf.com/content_cvpr_2018_workshops/w13/html/Yuan_Unsupervised_Image_Super-Resolution_CVPR_2018_paper.html.

Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., and Metaxas, D. N. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *CoRR*, abs/1612.03242, 2016. URL http://arxiv.org/abs/1612.03242.

Zhang, H., Goodfellow, I. J., Metaxas, D. N., and Odena, A. Self-attention generative adversarial networks. *CoRR*, abs/1805.08318, 2018. URL http://arxiv.org/abs/1805.08318.

Zhu, J., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017. URL http://arxiv.org/abs/1703.10593.