A PROOFS

A.1 PROOF OF LEMMA 2

Proof. In the limit when $h \to 0$, the Gaussian kernel converges to

$$K_h(t) = \begin{cases} 1/Z_h & \text{if } t = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the kernel $K_h(d_{\mathcal{X}}(x_i, x_k))$ will only have non-zero value when $x_i = x_k$, which implies that the kernelized mutual information will converge as follows:

$$\begin{aligned} \lim_{h\to 0} \hat{I}_\Gamma(X,Y) &= \lim_{h\to 0} \sum_{i,j} \Gamma_{ij} \log \frac{n^2 \cdot \sum_{k,l} \Gamma_{kl} K_h\left(d_{\mathcal{X}}(x_i,x_k)\right) K_h\left(d_{\mathcal{Y}}(y_j,y_l)\right)}{\sum_k K_h\left(d_{\mathcal{X}}(x_i,x_k)\right) \cdot \sum_l K_h\left(d_{\mathcal{Y}}(y_j,y_l)\right)}. \\ &= \sum_{i,j} \Gamma_{ij} \log \frac{n^2 \cdot \Gamma_{ij}/Z_h^2}{1/Z_h^2} \\ &= \sum_{i,j} \Gamma_{ij} \log \Gamma_{ij} + 2\log(n) \\ &= -H(\Gamma) + 2\log(n). \end{aligned}$$

□

A.2 PROJECTED GRADIENT DESCENT

The classic mirror descent iteration is written as:

$$x_{t+1} \leftarrow \arg\min_x \left\{ \tau \langle \nabla f(x_t), x \rangle + D(x \| x_t) \right\}.$$

When $D(y||x)$ is the KL divergence: $D_{\mathrm{KL}}(y||x) = \sum_i y_i \log \frac{y_i}{x_i}$, the update has the following form:

$$(x_{t+1})_i = e^{\log(x_t)_i - \tau \nabla f(x_t)} = (x_t)_i e^{-\tau \nabla f(x_t)}.$$

In our case, before the projection, the update reads

$$\Gamma'_{t+1} = \left(\Gamma_t \odot e^{-\tau(C - \lambda \nabla_{\Gamma_t} \hat{I}_{\Gamma_t}(X,Y) - \epsilon \nabla H(\Gamma_t))} \right).$$

Next, we solve the following projection w.r.t. KL metric:

$$\Gamma_{t+1} = \arg\min_{\Gamma \in \Pi(\mathbf{p},\mathbf{q})} D_{\mathrm{KL}}(\Gamma \| \Gamma'_{t+1}).$$

As Benamou et al. (2015) shows, the KL projection is equivalent to solving the entropic regularized optimal transport problem, which is usually refer to the sinkhorn distance (Cuturi, 2013). Following (Peyré et al., 2016), we set the stepsize $\tau = 1/\epsilon$ to simplify the iterations and reach the following update rule:

$$\Gamma_{t+1} \leftarrow \arg\min_{\Gamma \in \Pi(\mathbf{p},\mathbf{q})} \left\langle \Gamma, C - \lambda \nabla_\Gamma \hat{I}_{\Gamma_t}(X,Y) \right\rangle - \epsilon H(\Gamma).$$

B ADDITIONAL EXPERIMENTS

B.1 COLOR TRANSFER

Color transfer aims to transfer the colors of the target images into the source image. Optimal transport achieves this by treating pixels as points in the RGB space, and maps the source pixels to the target ones. Here, 500 pixels are sampled from each image to compute the OT, then the barycentric projection is applied to map all the source pixels to target. We compare fused InfoOT with standard OT, Sinkhorn distance (Cuturi, 2013), and linear mapping estimation (Perrot et al., 2016) and show the results in Figure 7. We can see that InfoOT produces a sharper results than the baselines while decently recovering the colors in the target image.

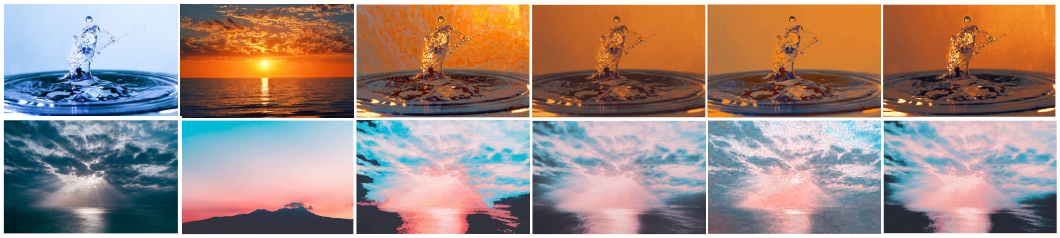

Figure 7: Color Transfer via Optimal Transport. Fused InfoOT produces sharper results while preserving the target color compared to the baselines.

	Supervision	EN-ES		EN-FR		EN-DE		EN-IT		EN-RU	
		→	←	→	←	→	←	→	←	→	←
PROCRUSTES	5K Words	81.2	82.3	81.2	82.2	73.6	71.9	76.3	75.5	51.7	63.7
Adv-NN	None	81.7	**83.3**	82.3	82.1	74.0	72.2	77.4	76.1	**52.4**	61.4
InvOT	None	81.3	81.8	82.9	81.6	73.8	71.1	77.7	77.7	41.7	55.4
InfoOT (h=0.55)	None	81.6	78.5	82.4	80.5	75.4	74.2	78.6	75.7	48.1	52.9
GW	None	**84.3**	83.2	**84.8**	**83.6**	**77.4**	**75.2**	**82.5**	**79.8**	52.0	**61.4**

Table 4: Cross-lingual Word Alignment. The InfoOT achieves comparable performance to GW, demonstrating its potential in recovering cross-lingual correspondence.

B.2 Word Embedding Alignment

Here, we explore the possibility of applying InfoOT for unsupervised word embedding alignment. We follow the setup in (Alvarez-Melis & Jaakkola, 2018), where the goal is to recover cross-lingual correspondences with word embedding in different languages. In this case, the pairwise distance between domains might not be meaningful, as the word embedding models are trained separately. Previous works suggest that cross-lingual word vector spaces are approximately isometric, which makes Gromov-Wasserstein an ideal choice due to its ability to align isometric spaces. Here, we treat GW as the oracle, and show that InfoOT can perform comparably to GW (Alvarez-Melis & Jaakkola, 2018) and other baselines such as InvOT (Alvarez-Melis et al., 2019), Adv-NN (Conneau et al., 2017), and supervised PROCRUSTES. We report the results on the dataset of Conneau et al. (2017) in Table 4, where both GW and InfoOT are trained with 12000 words and refined with Cross-Domain Similarity Scaling (CSLS) (Conneau et al., 2017). The entropy regularizer is 0.0001 and 0.02 for GW and InfoOT, respectively. We can see that InfoOT performs comparably with the baselines and GW, demonstrating its applicability in recovering cross-lingual correspondence.

B.3 Different Hyperparameter for InfoOT

Here, we report the performance of InfoOT with different weights for entropic regularizer and mutual information on domain adaptation. As Table 6 shows, Fused-InfoOT performs consistently well across different hyperparameter selections.

B.4 Additional Baseline: Unbalance OT

In this section, we additional include the results of unbalanced OT (UOT) Chizat et al. (2018); Frogner et al. (2015), which solves the following constrain optimization problem with generalized Sinkhorn-Knopp matrix scaling algorithm:

$$\min_{\Gamma \in \Pi(\mathbf{p},\mathbf{q})} \langle \Gamma, C \rangle + \epsilon_m D_{\mathrm{KL}}(\Gamma \mathbb{1}, \mathbf{p}) + \epsilon_m D_{\mathrm{KL}}(\Gamma^T \mathbb{1}, \mathbf{q}) - \epsilon H(\Gamma).$$

We show the best results of UOT in Table 5 by selecting ϵ and ϵ_m within $(1, 5, 10)$ and $(0.5, 1, 10)$. We can see that InfoOT still outperform UOT by a non-trivial margin.

	F-InfoOT*	UOT
C→D	**87.5±7.8**	81.9±11.9
C→W	**81.0±6.7**	77.3±6.4
C→A	**90.6±2.0**	87.8±3.5
D→W	**93.3±4.7**	93.3±5.7
D→A	**89.8±1.9**	87.8±3.2
D→C	**80.8±1.8**	78.8±2.7
W→D	89.4±11.8	**98.8±2.6**
W→C	74.4±3.7	**76.7±4.3**
W→A	**89.3±2.3**	80.1±3.3
A→D	**81.3±9.8**	76.3±8.2
A→W	**87.0±4.6**	69.3±8.6
A→C	**81.2±3.6**	77.4±3.7
AVG	**85.6±5.6**	82.1±8.3

Table 5: Unbalanced OT.

(λ, ϵ)	$(100, 1)$	$(100, 10)$	$(100, 20)$	$(10, 1)$	$(200, 1)$
C→D	87.5±7.8	86.3±8.7	85.0±8.9	87.5±7.8	86.9±8.0
C→W	81.0±6.7	86.7±7.7	88.0±5.9	80.0±6.1	81.7±7.2
C→A	90.6±2.0	90.5±2.1	90.7±2.1	89.4±2.4	90.6±2.0
D→W	93.3±4.7	92.7±6.0	91.3±5.3	93.3±5.7	94.0±4.4
D→A	89.8±1.9	89.9±2.1	89.6±1.9	89.6±1.7	89.8±1.5
D→C	80.8±1.8	81.2±1.9	81.5±1.6	80.7±1.8	80.6±1.8
W→D	89.4±11.8	86.9±10.4	83.8±11.5	91.9±12.2	90.0±11.9
W→C	74.4±3.7	74.2±3.7	74.0±3.4	74.2±4.6	74.4±3.8
W→A	89.3±2.3	89.3±2.0	89.3±2.0	86.4±2.8	89.3±2.3
A→D	81.3±9.8	80.6±9.5	82.5±11.7	81.9±10.4	82.5±9.2
A→W	87.0±4.6	83.3±6.3	83.0±6.0	83.8±6.5	87.0±4.6
A→C	81.2±3.6	80.8±4.0	80.2±3.9	81.2±3.3	82.2±2.7
AVG	85.6±5.6	85.2±5.3	84.9±5.1	85.0±5.6	85.7±5.5

Table 6: InfoOT with different hyperparameters. We test the Fused-InfoOT with conditional projection by varying the regularizer weights (λ, ϵ). Note that Table 1 in the main paper shows the results of $(\lambda = 100, \epsilon = 1)$.

	1-NN	5-NN	10-NN	20-NN	Linear
OT	71.0±8.8	77.0±6.4	78.3±4.8	77.5±6.8	77.8±8.6
Sinkhorn	72.4±7.3	76.0±5.3	76.3±4.0	75.2±6.3	76.7±9.8
GL-OT	78.9±7.3	80.7±5.3	80.5±4.3	78.2±7.1	78.1±9.7
FGW	71.0±8.8	76.9±6.5	78.3±4.8	77.5±6.8	77.5±7.9
Linear	70.5±8.4	75.9±6.2	77.4±5.4	77.5±7.1	76.7±7.5
F-InfoOT	80.6±5.7	81.4±5.8	79.7±5.0	76.4±7.1	**82.9±7.0**
F-InfoOT*	**85.5±5.6**	**85.4±5.5**	**85.4±5.5**	**81.7±7.2**	81.4±5.3

Table 7: Results beyond 1-NN. We evaluate the performance with k-NN classifiers and linear classifiers.

	GFK	CORAL	SCA	JDA	TJM	DDC	DAN	MEDA	F-InfoOT*
C→D	86.6	84.7	87.9	89.8	84.7	88.8	89.3	91.1	87.9
C→W	77.6	80.0	85.4	85.1	81.4	85.4	90.6	95.6	85.8
C→A	88.2	92.0	89.5	89.6	88.8	91.9	92.0	93.4	91.1
D→W	99.3	99.3	98.6	99.7	99.3	98.2	98.5	97.6	97.3
D→A	76.3	85.5	90.0	91.7	90.3	89.5	90.0	93.2	91.3
D→C	71.4	76.8	78.1	85.5	83.8	81.1	80.3	87.5	82.9
W→D	100	100	100	100	100	100	100	99.4	96.2
W→C	69.8	75.5	74.8	84.8	83.0	78.0	81.2	93.2	80.3
W→A	76.8	81.2	86.1	90.3	84.6	84.9	92.1	99.4	90.0
A→D	82.2	84.1	85.4	80.3	76.4	89.0	91.7	88.1	81.5
A→W	70.9	74.6	75.9	78.3	71.9	86.1	91.8	88.1	85.4
A→C	79.2	83.2	78.8	83.6	84.3	85.0	84.1	87.4	82.5
AVG	81.5	84.7	85.9	88.2	86.0	88.2	90.1	92.8	87.7

Table 8: Baselines beyond OT.

B.5 Experiments beyond 1-NN Classifier

We report the performances of InfoOT and baselines with general k-NN classifiers and linear SVM classifiers in Table 7. We can see that fused-InfoOT consistently outperforms the baselines beyond 1-NN classifiers on Office-Caltech domain adaptation benchmark. In addition, compared to the baselines, the performance of InfoOT is more robust to the choice of the number of neighbors k.

B.6 Baselines beyond Optimal Transport

We compare InfoOT with the following non-OT baselines: Geodesic Flow Kernel (GFK) (Gong et al., 2012), CORrelation Alignment (CORAL) (Sun et al., 2016), Scatter Component Analysis (SCA) (Ghifary et al., 2016), Joint distribution alignment (JDA) (Long et al., 2013), Transfer Joint

Matching (TJM) (Long et al., 2014a), Deep Domain Confusion (DDC) (Tzeng et al., 2014), Deep Adaptation Network (DAN) (Long et al., 2014b), and Manifold Embedded Distribution Alignment (MEDA) (Wang et al., 2018). For fair comparison, we report the performance of Fused-InfoOT calculated with full source and target dataset instead of the 10-fold setting in the main context. As Table 8 shows, InfoOT performs comparably to many baselines without training or finetuning neural networks.

C LIMITATIONS

While we have illustrated successful applications of InfoOT, there are limitations. One could expect InfoOT to perform worse when the geometry of input spaces provides little information. In particular, for raw inputs such as image datasets, InfoOT would not perform well without pre-extracted features. It is also non-trivial to directly apply InfoOT to very large-scale problems with millions of data points. Computational-efficient extensions such as mini-batch optimal transport (Nguyen et al., 2022) should be considered to apply InfoOT to large-scale datasets.