Immiscible Color Flows in Optimal Transport Networks for Image Classification

Alessandro Lonardi,[1, *] Diego Baptista,[1, †] and Caterina De Bacco[1, ‡]

[1]*Max Planck Institute for Intelligent Systems, Cyber Valley, Tübingen 72076, Germany*

In classification tasks, it is crucial to meaningfully exploit information contained in data. Here, we propose a physics-inspired dynamical system that adapts Optimal Transport principles to effectively leverage color distributions of images. Our dynamics regulates immiscible fluxes of colors traveling on a network built from images. Instead of aggregating colors together, it treats them as different commodities that interact with a shared capacity on edges. Our method outperforms competitor algorithms on image classification tasks in datasets where color information matters.

Optimal Transport (OT) is a powerful method for computing the distance between two data distributions. This problem has a cross-disciplinary domain of applications, ranging from logistic and route optimization [1–3], to biology [4, 5] and computer vision [6–9], among others. Within this broad variety of problems, OT is largely utilized in machine learning [10], and deployed for solving classification tasks, where the goal is to optimally match discrete distributions that are typically learned from data. Relevant usage examples are also found in multiple fields of physics, as in protein fold recognition [11], stochastic thermodynamics [12], routing in multi-layer networks [13] or general relativity [14].

A prominent application is image classification [15–20], where the goal is to measure the similarity between two images. OT solves this problem by interpreting image pairs as two discrete distributions and then assesses their similarity via the Wasserstein (W_1) distance [21], a measure obtained by minimizing the cost needed to transform one distribution into the other.

Much research effort has been put in speeding up algorithms to calculate this quantity [11, 16, 17, 22, 23]. However, all these methods overlook the potential of using effectively image colors. As a result, practitioners have access to increasingly efficient algorithms, but that do not necessarily improve accuracy in predictions, as we lack a framework that fully exploits the richness of the input information. Colored images, originally encoded into three-dimensional histograms—with one dimension per color channel—are often compressed into lower dimensional data using feature extraction algorithms [9, 20]. Here, we propose a different approach that maps the three distinct color histograms to multicommodity flows transported in a network built using images' pixels. We combine recent developments of OT with physics insights of capacitated network models [1, 5, 24–27] to treat colors as mass of different types that flows through the edges of a network. Different flows are coupled together with a shared conductivity, and minimize a unique cost function. This setup is reminiscent of the distinction between modeling the flow of one substance, e.g. water, and modeling flows of multiple substances that do not mix, e.g. immiscible fluids, that share the same network infrastructure.

By virtue of this multicommodity treatment we achieve stronger classification performance than state-of-the-art OT-based algorithms in real datasets where color information matters.

Unicommodity Optimal Transport.—Given two m, n-dimensional probability vectors g, h, and a positive-valued ground cost matrix C, the goal of a standard—unicommodity—OT problem is to find an Optimal Transport path $P^\star$ satisfying the conservation constraints $\sum_j P_{ij} = g_i \ \forall i$ and $\sum_i P_{ij} = h_j \ \forall j$, while minimizing $J(g, h) = \sum_{ij} P_{ij} C_{ij}$. Entries $P_{ij}^\star$ can be interpreted as the mass transported from g_i to h_j when paying a cost C_{ij}, while $J^\star$, i.e. J evaluated at $P^\star$, encodes the minimum effort needed to transport g to h. Notably, if all entries C_{ij} are distances between i and j, then $J^\star$ is the W_1 distance between g and h (see [21] for a standard proof and [9] for derivations focusing on the discrete case).

Multicommodity Optimal Transport.—We adapt unicommodity OT to account for M different histograms that are coupled together. In our application, this amounts to considering $M = 3$ color channels, the *commodities*, indexed with $a = 1, \ldots, M$. We define g^a, h^a as m, n-dimensional probability vectors of mass of type a. More compactly, we define the matrix G with entries $G_{ia} = g_i^a$ (resp. H for h), each containing the intensity of color channel a in pixel i of the first (resp. second) image. Conservation of mass must hold for each commodity index a, i.e. $\sum_i g_i^a = \sum_j h_j^a$.

Moreover, we define the set $\Pi(G, H)$, containing $(m \times n \times M)$-dimensional tensors P with entries P_{ij}^a, being transport paths between g^a and h^a. We enforce interaction between transport paths for different commodities by introducing a *shared cost* C. This allows us to define the multicommodity transport cost $J_\Gamma(G, H) = \sum_{ij} ||P_{ij}||_2^\Gamma C_{ij}$, where $||P_{ij}||_2$ is the 2-norm of the vector $P_{ij} = (P_{ij}^1, \ldots, P_{ij}^M)$ and $0 < \Gamma < 4/3$ is a regularization parameter. We can thus formulate its corresponding multicommodity OT problem as that of finding a tensor $P^\star$ solution of:

$$J_\Gamma^\star(G, H) = \min_{P \in \Pi(G,H)} J_\Gamma(G, H). \qquad (1)$$

Notice that for $M = 1$ and $\Gamma = 1$, we recover the standard unicommodity OT setup. Our choice of using

the 2-norm over a in J_Γ is motivated by recent works [2, 28, 29] where it is shown that Eq. (1) corresponds to Joule's law, while transport paths follow Kirchhoff's law enforcing conservation of mass.

Optimal Transport network on images.—To adapt this formalism to images, we introduce an auxiliary bipartite network $K_{m,n}(V_1, V_2, E_{12})$, being the first building block of the network where the OT problem with input matrices G, H—representing two images—is solved. The sets of nodes V_1, V_2 are the pixels of Image 1 and Image 2, respectively. The set of edges E_{12} contains a subset of all pixels' pairs between the two images, as detailed below.

We consider the cost of an edge (i, j) as $C_{ij}(\theta, \tau) = \min\{(1-\theta)\|\mathrm{v}_i - \mathrm{v}_j\|_2 + \theta\|G_i - H_j\|_1, \tau\}$, where the vector $\mathrm{v}_i = (x_i, y_i)$ contains the coordinates of pixel i of Image 1 (similarly v_j for Image 2), $0 \le \theta \le 1$ is a convex combination weight, and $\tau > 0$ is a trimming threshold. This definition of cost interpolates between the Euclidean distance between pixels and the difference in their color intensities, following the intuition in [9, 20]. The relative influence of these two contributions is tuned by θ, an hyperparameter given in input.

The parameter τ is introduced following [19, 20], with the scope of removing all edges with cost $C_{ij}(\theta, \tau) = \tau$ and substituting them with $m+n$ transshipment edges $e \in E'$, each of cost $\tau/2$, connected to one unique auxiliary vertex u_1. Thresholding the cost decreases significantly the computational complexity of OT, making it linear with the number of nodes $|V_1| + |V_2| + 2 = m + n + 2$ (see Supplemental Material [30]).
OT is solved by injecting in nodes $i \in V_1$ the color mass specified by G and extracting it in nodes $j \in V_2$, as specified by H. This is done by transporting mass using either (i) an edge in E_{12} or (ii) a transshipment one in E'.

Furthermore, we relax conservation of mass by allowing $\sum_i G_{ia} \ne \sum_j H_{ja}$, and assigning mass contributions $m^a = \sum_j H_{ja} - \sum_i G_{ia}$ to a second auxiliary node, u_2. We connect it to the network with n additional transshipment edges $e \in E'$, each penalizing the total cost by $c = \max_{ij} C_{ij}/2$. This construction improves classification when the histograms' total mass largely differs [19]. Intuitively, this can happen when comparing "darker" images against "brighter" ones. More precisely, when entries of g^a and h^a are further apart in the RGB color space.

Overall we obtain a network K with nodes $V = V_1 \cup V_2 \cup \{u_1, u_2\}$ and edges $E = E_{12} \cup E'$, i.e. the original bipartite graph $K_{m,n}$, together with the auxiliary transshipment links and nodes. Note that in its entirety the system is isolated, i.e. the total mass is conserved, see Fig. 1 for a representation of the OT setup and the Supplemental Material [30] for a detailed description.

Optimizing immiscible color flows: the dynamics.—We solve the OT problem by proposing the following ODEs
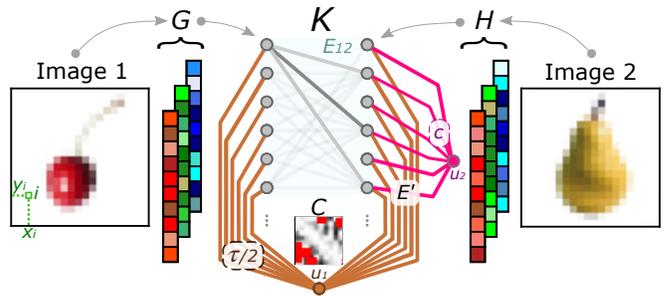
FIG. 1. Network for multicommodity OT. Images are RGB matrices, G and H, that flow into K. The graph is made of $n + m + 2$ nodes, one per pixel, and two auxiliary vertices. Edges in E_{12} (gray) are trimmed with τ (C_{ij} is highlighted in red if larger than τ). Transshipment, and auxiliary links to relax mass conservation, in E' are in brown and magenta.

controlling mass transportation:

$$\sum_{j \in \partial i} L_{ij}[x]\phi_j^a = S_i^a \qquad \forall i \in V, a = 1, \ldots, M \tag{2}$$

$$\frac{dx_e}{dt} = x_e^\beta \frac{\|\phi_i - \phi_j\|_2^2}{C_e^2} - x_e \quad \forall e = (i, j) \in E, \tag{3}$$

which constitute the pivotal equations of our model. Here we introduce the *shared conductivities* $x_e \ge 0$, and define $S_i^a = G_{ia} - H_{ia}$, taking values $S_{u_1}^a = 0$ and $S_{u_2}^a = m^a$. With $L_{ij}[x] = \sum_e (x_e/C_e) B_{ie} B_{je}$ we denote the weighted Laplacian of K, where B is its signed incidence matrix; ∂i is the neighborhood of node i. Lastly, ϕ_i^a are scalar potentials acting on nodes, for a given commodity a. Least-square solutions of Eq. (2) are $\phi_i^a[x] = \sum_j L_{ij}^\dagger[x]S_i^a$, where $\dagger$ denotes the Moore-Penrose inverse.

The feedback mechanism of Eq. (3) allows to allocate multicommodity fluxes of color of type a (i.e. transport paths) on edges $e = (i, j)$ by defining $P_e^a(t) = x_e(t)(\phi_i^a[x(t)] - \phi_j^a[x(t)])/C_e$. Crucially, this allocation is governed by *one unique conductivity* for all commodities, whose dynamics depends on the 2-norm of differences of potentials over a. Instead, transport paths are commodity-dependent because of the potentials ϕ. The critical exponent $0 < \beta < 2$ [$\Gamma = 2(2 - \beta)/(3 - \beta)$] aggregates paths with a principle of economy of scale if $1 < \beta < 2$. It dilutes them along the network otherwise, with the goal of reducing traffic congestion.
In the case of only one commodity ($M = 1$), variants of this dynamics have been used to model transport optimization in various physical systems [1, 5, 25–27].

The salient result of this construction is that asymptotic trajectories of Eqs. (2)-(3) are equivalent to minimizers of Eq. (1), i.e. $\lim_{t \to +\infty} P(t) = P^\star$ (see Supplemental Material [30] for derivations following [28, 29]). Therefore, numerically integrating our dynamics solves the multicommodity OT problem.

Classification task.—We provide empirical evidence that our multicommodity dynamics outperforms competing OT algorithms on classification tasks. Given two images, we use the OT optimal cost as a measure of similarity between them, and perform supervised classification with a k-nearest neighbor (k-NN) classifier.

We compare classification accuracy of our model against: (i) Sinkhorn algorithm [16, 31]; (ii) a unicommodity dynamics executed on grayscale images, i.e., with color information compressed in one single commodity ($M = 1$); (iii) Sinkhorn algorithm on grayscale images. All methods are tested on two datasets: the Jena Flowers 30 Dataset (JF30) [32], and the Fruit Dataset (FD) of [33]. The first consists of 1,479 images of 30 wild-flowering angiosperms (flowers). Flowers are labelled with their species, inferring them is the goal of the classification task. The second dataset contains 15 fruit types and 163 images, here we want to classify fruit types. Parameters of the OT problem setup (θ and τ), as well as regularization ones (β and ε, which enforces the entropic barrier in Sinkhorn algorithm [16]), have been cross-validated for both datasets. All methods are then tested in their optimal configurations (see Supplemental Material [30] for implementation details).

Classification results are shown in Table I. In all cases, leveraging colors leads to higher accuracy (about 8% increase) with respect to classification performed using grayscale images. This signals that, in the datasets under consideration, color information is a relevant feature for differentiating image samples. Remarkably, we get a similar increase in performance (about 7-8%) on both colored datasets, when comparing our multicommodity dynamics against Sinkhorn. As the two algorithms use the same (colored) input, we can attribute this increment to an effective usage of the color that our approach is capable of.

In addition, analyzing results in more detail, we first observe that on JF30 all methods perform best when $\theta = 0.25$, i.e., 25% percent of the information used to build C comes from colors. This trend does not recur on FD, where both dynamics favor $\theta = 0$ (Euclidean C). Hence, our model is able to leverage color information via the multicommodity OT dynamical formulation.

Second, on JF30 both dynamics perform best with $\tau = 0.125$, contrarily to Sinkhorn-based methods that prefer $\tau = 0.05$. Thus, Sinkhorn's classification accuracy is negatively affected both by low τ—many edges of the transport network are cut, and by τ large—noisy color information is used to build C. We do not observe this behavior in our model, where trimming less edges is advantageous. All optimal values of τ are lower on FD, since color distributions of this dataset are naturally light-tailed (See Supplemental Material [30]).

Lastly, we investigate the interplay between θ and β. We notice that $\theta = 0$ (FD) corresponds to higher $\beta = 1.5$. Instead, for larger $\theta = 0.25$ (JF30), the model prefers lower β ($\beta = 1, 1.25$ for the multicommodity and the unicommodity dynamics, respectively). In the former case ($\theta = 0$, C_{ij} is the Euclidean distance), the cost is equal to zero for pixels with the same locations. Thus, consolidation of transport paths—large β—is favored on cheap links. Instead, increasing θ leads to more edges with comparable costs as colors distribute smoothly over images. In this second scenario, better performance is achieved with distributed transport paths, i.e. lower β (See Supplemental Material [30]).

The impact of colors.—To further assess the importance of leveraging color information we conduct three different experiments.

In Experiment 1, we focus on a single image taken from the test set of FD, and we plot the landscape of optimal costs $J_\Gamma^\star$ when comparing it against the train set. Results for the multicommodity dynamics ($M = 3$) and for the unicommodity dynamics ($M = 1$) on grayscale images are in Fig. 2. Here, we highlight the 5 lowest values of the cost, and mark them in green if they correspond to correctly classified train samples, and in red otherwise. Both dynamics classify correctly 3 samples out if 5. However, the multicommodity dynamics clusters them at the bottom of the cost landscape, thus improving classification.

We further mark this tendency with Experiment 2. Here, we perform classification on a subset of FD composed of images belonging to three classes: red apples, orange apricots, yellow melons. These fruits have similar shapes, so their color plays a crucial role in the classification process. The test set is made of three random samples each drawn from one of these classes (top row of the rightmost panel), the train set contains the remaining instances of the classes. Results are in Fig. 2. First, we plot the cost landscape $J_\Gamma^\star$ for the train set, and draw in red,

	Algorithm	Hyperparameters					Class. accuracy
		θ	τ	β	ε	k	[%] ($\uparrow$)
JF30	Multicommodity	0.25	0.125	1	—	1	**62.2**
	Sinkhorn RGB	0.25	0.05	—	100	1	58.4
	Sinkhorn GS	0.25	0.05	—	500	1	54.3
	Unicommodity	0.25	0.125	1.25	—	1	53.6
FD	Multicommodity	0	0.04	1.5	—	2	**75.0**
	Sinkhorn RGB	0.5	0.06	—	750	1	69.6
	Unicommodity	0	0.06	1.5	—	5	64.3
	Sinkhorn GS	0.25	0.06	—	500	4	60.7

TABLE I. Classification task results. With Multicommodity, Sinkhorn RGB, Unicommodity and Sinkhorn GS we label methods on colored images (first two), and grayscale ones (second two). Optimal parameters in the central columns are selected with a 4-fold cross-validation; k is the number of nearest neighbors used in the classifier. The rightmost column shows the fraction (in percentage) of correctly classified images. Results are ordered by performance, in bold we highlight the best ones.
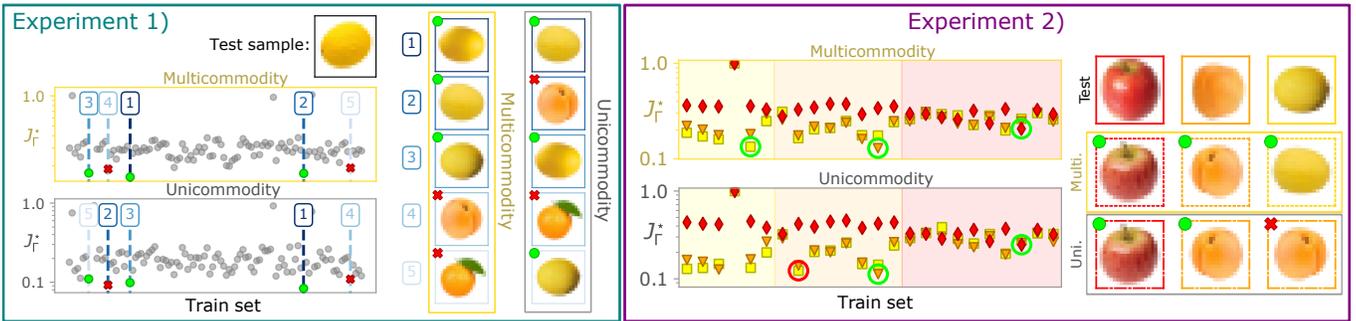
FIG. 2. Evaluating the effect of colors. Experiment 1: Top black framed image is that to be classified. Predictions given by the multicommodity and the unicommodity dynamics (those with lower $J_\Gamma^\star$) are shown in the right side of the panel, and are displayed in sorted fashion from worst to best (from bottom to top). Experiment 2: Top right samples are the three test images to be classified. Middle and bottom rows are predictions given by the two dynamics. Markers, backgrounds and test images shared color code: red for apples, orange for apricots, yellow for melons. In both panels: green circles and red crosses are used to highlight classified and misclassified images, respectively. All algorithms are executed with their optimal configurations in Table I.

orange, and yellow values of $J_\Gamma^\star$ correspondent to samples which are compared against the test apple, apricot, and melon, respectively. We also sort the train samples so that they are grouped in three regions (highlighted by the background color in Fig. 2), correspondent to train melons, apricots, and apples. With this construction, if the minimum cost among the yellow markers falls in the yellow region, it will correspond to a correctly classified sample (resp. for orange and red). We further mark the yellow, orange, and red minima in green if test and train labels correspond, i.e. markers' and backround's colors are the same, and in red otherwise. Train and test samples are also in Fig. 2. The multicommodity dynamics correctly labels each test image. In contrast, the unicommodity dynamics fails at this task, labeling a melon as an apricot.

Lastly, in Experiment 3 we consider a setup where fruits' shapes crucially influence classification. In particular, we select as test sample a cherry, whose form is arguably clearly distinguishable from that of many other fruits in the dataset. One can expect that comparing it against the train set of FD will result in having *both* unicommodity and multicommodity dynamics able to assign low $J_\Gamma^\star$ to train cherries, and higher costs to other fruits. This intuition is confirmed by results in Fig. 3. Here train cherries (in green) strongly cluster in the lower portion of the cost landscape, whereas all the other fruits have higher costs. In Fig. 3 we also plot some of the correctly classified train samples.

These results suggest that when color information is negligible compared to another type of information (e.g. shape), unicommodity and multicommodity formulations perform similarly. In light of this, we reinforce the claim that our multicommodity formulation can boost classification in contexts where color information does matter.

Conclusions.—We propose a multicommodity OT formulation for effectively using the color information to im-
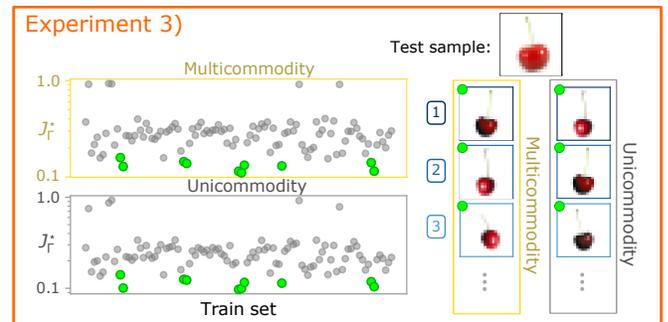
FIG. 3. Evaluating the importance of colors: when shapes matter most. Experiment 3: Top black framed image is that to be classified. The best 3 (out of 10) predictions returned by the two dynamics are shown on the right. We mark with green circles training samples belonging to the same class of the test image. All algorithms are executed with their optimal configurations in Table I.

prove image classification. We model colors as immiscible flows traveling on a capacitated network, and propose equations for its dynamics, with the goal of optimizing flow distribution on edges. Color flows are regulated by a shared conductivity, and minimize a unique cost function. Thresholding the ground cost as in [19, 20] makes our model computationally efficient.

We outperform Sinkhorn algorithm on two datasets where color matters. Our model also assigns lower cost to correctly classified images than its unicommodity counterpart, and it is more robust on datasets where items have similar shape, and thus color information is distinctly relevant. We note that, for some datasets, color information may not matter, as when another type of information (e.g. shape) has stronger discriminative power. However, while we focused here on different color channels as the different commodities in our formulation, the ideas of this paper can be extended to scenarios where

other relevant information can be distinguished into different types.

To facilitate practitioners using our algorithms, we make our code publicly available [34].

---

[*] Equal contribution; alessandro.lonardi@tuebingen.mpg.de

[†] Equal contribution; diego.theuerkauf@tuebingen.mpg.de

[‡] caterina.debacco@tuebingen.mpg.de

[1] F. Kaiser, H. Ronellenfitsch, and D. Witthaut, Discontinuous transition to loop formation in optimal supply networks, Nature communications **11**, 1 (2020).

[2] A. Lonardi, M. Putti, and C. De Bacco, Multicommodity routing optimization for engineering networks (2021), arXiv:2110.06171 [physics.soc-ph].

[3] A. Lonardi, E. Facca, M. Putti, and C. De Bacco, Infrastructure adaptation and emergence of loops in network routing with time-dependent loads (2021), arXiv:2112.10620 [physics.soc-ph].

[4] P. Demetci, R. Santorella, B. Sandstede, W. S. Noble, and R. Singh, Gromov-Wasserstein optimal transport to align single-cell multi-omics data, bioRxiv 10.1101/2020.04.28.066787 (2020).

[5] E. Katifori, G. J. Szöllősi, and M. O. Magnasco, Damage and Fluctuations Induce Loops in Optimal Transport Networks, Phys. Rev. Lett. **104**, 048704 (2010).

[6] M. Werman, S. Peleg, and A. Rosenfeld, A Distance Metric for Multidimensional Histograms, Computer Vision, Graphics, and Image Processing **32**, 328 (1985).

[7] S. Peleg, M. Werman, and H. Rom, A unified approach to the change of resolution: space and gray-level, IEEE Transactions on Pattern Analysis and Machine Intelligence **11**, 739 (1989).

[8] Y. Rubner, C. Tomasi, and L. J. Guibas, A metric for distributions with applications to image databases, in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)* (1998) pp. 59–66.

[9] Y. Rubner, C. Tomasi, and L. J. Guibas, The Earth Mover's Distance as a Metric for Image Retrieval, International Journal of Computer Vision **40**, 99 (2000).

[10] G. Peyré and M. Cuturi, Computational Optimal Transport: With Applications to Data Science, Foundations and Trends® in Machine Learning **11**, 355 (2019).

[11] P. Koehl, M. Delarue, and H. Orland, Optimal transport at finite temperature, Phys. Rev. E **100**, 013310 (2019).

[12] E. Aurell, C. Mejía-Monasterio, and P. Muratore-Ginanneschi, Optimal Protocols and Optimal Transport in Stochastic Thermodynamics, Phys. Rev. Lett. **106**, 250601 (2011).

[13] A. A. Ibrahim, A. Lonardi, and C. D. Bacco, Optimal Transport in Multilayer Networks for Traffic Flow Optimization, Algorithms **14**, 10.3390/a14070189 (2021).

[14] A. Mondino and S. Suhr, An optimal transport formulation of the Einstein equations of general relativity, Journal of the European Mathematical Society DOI: 10.4171/JEMS/1188 (2022).

[15] K. Grauman and T. Darrell, Fast contour matching using approximate earth mover's distance, in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*, Vol. 1 (2004) pp. I–I.

[16] M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, in *Advances in Neural Information Processing Systems*, Vol. 26 (Curran Associates, Inc., 2013) pp. 2292–2300.

[17] P. Koehl, M. Delarue, and H. Orland, Statistical Physics Approach to the Optimal Transport Problem, Phys. Rev. Lett. **123**, 040603 (2019).

[18] M. Thorpe, S. Park, S. Kolouri, G. K. Rohde, and D. Slepčev, A transportation L^p Distance for Signal Analysis, Journal of mathematical imaging and vision **59**, 187 (2017).

[19] O. Pele and M. Werman, A Linear Time Histogram Metric for Improved SIFT Matching, in *Computer Vision – ECCV 2008* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2008) pp. 495–508.

[20] O. Pele and M. Werman, Fast and robust Earth Mover's Distances, in *2009 IEEE 12th International Conference on Computer Vision* (2009) pp. 460–467.

[21] C. Villani, *Optimal transport: Old and New*, Vol. 338 (Springer, 2009).

[22] T. Lin, N. Ho, and M. Jordan, On Efficient Optimal Transport: An Analysis of Greedy and Accelerated Mirror Descent Algorithms, in *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 97 (PMLR, 2019) pp. 3982–3991.

[23] P. Dvurechensky, A. Gasnikov, and A. Kroshnin, Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm, in *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 80 (PMLR, 2018) pp. 1367–1376.

[24] J. R. Banavar, F. Colaiori, A. Flammini, A. Maritan, and A. Rinaldo, Topology of the Fittest Transportation Network, Phys. Rev. Lett. **84**, 4745 (2000).

[25] H. Ronellenfitsch and E. Katifori, Global Optimization, Local Adaptation, and the Role of Growth in Distribution Networks, Phys. Rev. Lett. **117**, 138301 (2016).

[26] D. Hu and D. Cai, Adaptation and Optimization of Biological Transport Networks, Phys. Rev. Lett. **111**, 138701 (2013).

[27] F. Corson, Fluctuations and Redundancy in Optimal Transport Networks, Phys. Rev. Lett. **104**, 048703 (2010).

[28] A. Lonardi, E. Facca, M. Putti, and C. De Bacco, Designing optimal networks for multicommodity transport problem, Phys. Rev. Research **3**, 043010 (2021).

[29] V. Bonifaci, E. Facca, F. Folz, A. Karrenbauer, P. Kolev, K. Mehlhorn, G. Morigi, G. Shahkarami, and Q. Vermande, Physarum-inspired multi-commodity flow dynamics, Theoretical Computer Science https://doi.org/10.1016/j.tcs.2022.02.001 (2022).

[30] See Supplemental Material at [URL will be inserted by publisher].

[31] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet,

A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer, POT: Python Optimal Transport, Journal of Machine Learning Research **22**, 1 (2021).

[32] M. Seeland, M. Rzanny, N. Alaqraa, J. Wäldchen, and P. Mäder, Jena Flowers 30 Dataset (2017).

[33] P. A. Macanhã, D. M. Eler, R. E. Garcia, and W. E. M. Junior, Handwritten feature descriptor methods applied to fruit classification, in *Information Technology - New Generations* (Springer International Publishing, Cham, 2018) pp. 699–705.

[34] https://github.com/aleable/MODI.

[35] R. N. Shepard, Toward a Universal Law of Generalization for Psychological Science, Science **237**, 1317 (1987).

[36] V. Bonifaci, K. Mehlhorn, and G. Varma, Physarum can compute shortest paths, Journal of Theoretical Biology **309**, 121 (2012).

[37] C.A. Poynton, Frequently Asked Questions about Color.

[38] R. Sinkhorn, A relationship between arbitrary positive matrices and doubly stochastic matrices, The annals of mathematical statistics **35**, 876 (1964).

[39] P. Knopp and R. Sinkhorn, Concerning nonnegative matrices and doubly stochastic matrices., Pacific Journal of Mathematics **21**, 343 (1967).

[40] A. Genevay, M. Cuturi, G. Peyré, and F. Bach, Stochastic optimization for large-scale optimal transport, in *Advances in Neural Information Processing Systems*, Vol. 29, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., 2016).

[41] J. Altschuler, J. Niles-Weed, and P. Rigollet, Near-linear time approximation algorithms for optimal transport via sinkhorn iteration, in *Advances in Neural Information Processing Systems*, Vol. 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017).

[42] E. Facca and M. Benzi, Fast Iterative Solution of the Optimal Transport Problem on Graphs, SIAM Journal on Scientific Computing **43**, A2295 (2021).

[43] B. Schmitzer, Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems, SIAM Journal on Scientific Computing **41**, A1443 (2019).

[44] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, Scaling algorithms for unbalanced optimal transport problems, Mathematics of Computation **87**, 2563 (2018).

Immiscible Color Flows in Optimal Transport Networks for Image Classification: Supplemental Material (SM)

Alessandro Lonardi,[1, *] Diego Baptista,[1, †] and Caterina De Bacco[1, ‡]

[1]*Max Planck Institute for Intelligent Systems, Cyber Valley, Tübingen 72076, Germany*

CONSTRUCTION OF THE NETWORK

Here, we extensively explain the procedure used for constructing the transport network K. All the essential steps are schematically drawn in Fig. S1, and are as follows.

Step 1. Initially, a pair of images (Image 1, Image 2), is given as a couple of multidimensional arrays of dimensions $(w_i, h_i, M = 3)$, with $i = 1, 2$. We denote with w images' widths and with h their heights. The third dimension has size $M = 3$, and corresponds to the three RGB color channels. The color channels are flattened to obtain the tensors G and H, the first for Image 1, and the second for Image 2. In detail, each channel is vectorized to have dimension $m \times 1$, with $m = w_1 \cdot h_1$, for g^a (resp. $n \times 1$, with $n = w_2 \cdot h_2$, for h^a), which are inflows and outflows of our multicommodity dynamics. In this way, the tensors G and H, which are obtained stacking g^a and h^a, have size $m \times M$ and $n \times M$. Entries of G and H are in standard RGB encoding, hence they are integers ranging from 0 to 255.

To obtain the transport network K, we first generate a complete bipartite graph between $m + n = |V_1| + |V_2|$ nodes, the first $m = |V_1|$ are the pixels of Image 1, and the other $n = |V_2|$ are those of Image 2. We then assign a cost to each edge of this graph using both information given by pixels' locations, and by images' colors. In particular, we define:

$$C_e(\theta) = (1 - \theta)Y_e + \theta X_e \quad \forall e = (i, j) \tag{S1}$$

$$Y_{e=(i,j)} = ||v_i - v_j||_2 = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{S2}$$

$$X_{e=(i,j)} = ||G_i - H_j||_1 = \sum_{a=1}^{M=3} |G_{ia} - H_{ja}| \tag{S3}$$

for each i pixel of Image 1, and j pixel of Image 2. Terms Y_e in Eq. (S2) contain the Euclidean distance between any pair of pixels, whose horizontal and vertical coordinates are stored in vectors $v = (x, y)$. Instead, X_e contributes with colors to edges' costs. We model the effect of colors taking, in Eq. (S3), the 1-norm between arrays G_i, H_j, containing the RGB intensities in i (pixel of Image 1) and j (pixel of Image 2). Both X_e and Y_e have been opportunely rescaled in the range $[0, 1]$. Lastly, we use the scalar parameter $0 \leq \theta \leq 1$ to weigh Y_e and X_e in a convex combination, in Eq. (S1).

Step 2. Once Step 1 is complete, and a cost C_e is assigned to each edge of the complete bipartite graph between the two images, we implement a trimming procedure similar to that of [19, 20] to cut highly expensive links. In particular, we trim all edges e that have cost $C_e > \tau$, where $\tau > 0$ is a threshold fixed a priori. The links between V_1 and V_2 that do not get cut make up the set E_{12}. We then add a first transshipment node, u_1, to the network, and connect it with $m + n$ links to the sets V_1 and V_2. Each transshipment link is assigned a fixed cost $C_e = \tau/2$. This implies that one needs to pay a total cost of τ to transport mass from a node of Image 1 (in V_1) to one of Image 2 (in V_2), when traversing transshipment links.

There are several benefits in thresholding for the cost: (i) from a purely intuitive standpoint, humans perceive distances as saturated distances [35]; (ii) many natural color distributions are noisy and heavy-tailed, thus thresholding permits to assign a fixed cost to outliars; (iii) thresholded distances induce a W_1 distance between distributions in standard unicommodity OT problems [20]. More practically, thresholding improves accuracy and speed of OT [20] (see also the Computational Cost Section in this SM).

Step 3. The last step required to obtain K is the introduction of a second auxiliary node, u_2, together with its edges, to relax mass balance. In detail, in a standard OT setting $\sum_i G_{ia} = \sum_j H_{ja} = \Lambda^a > 0$ holds $\forall a = 1, \ldots, M$, i.e., two histograms to be transported belong to the same simplex of mass $\Lambda^a > 0$. We relax this constraint permitting $\sum_i G_{ia} \neq \sum_j H_{ja}$ and penalizing Eq. (1). Particularly, we use a similar relaxation of that in [20],

which we generalize to the multicommodity setup:

$$J_\Gamma^\star(G,H) = \min_{P\in\Pi(G,H)} J_\Gamma(G,H) \quad \overset{\text{Relaxation}}{\longrightarrow} \quad \min_P \left\{ J_\Gamma(G,H) + \alpha \sum_a \left| \sum_j H_{ja} - \sum_i G_{ia} \right| \max_{e\in E_{12}} C_e \right\}. \tag{S4}$$

The intuition of Eq. (S4) is that the OT problem is penalized proportionally to the net difference between the inflowing and the outflowing mass. Hence, two images whose colors strongly differ return a higher cost and, in a supervised classification task, are less likely to be assigned the same label. We fix $\alpha = 1/2$ as in [19].

This penalization can be translated to the transport network with the addition of n links, costing $c = \alpha \max_{e\in E_{12}} C_e = \max_{e\in E_{12}} C_e/2$, connected to u_2. The excess of mass $m^a = \sum_j H_{ja} - \sum_i G_{ia}$ given by each commodity is injected u_2 to guarantee that the whole system is isolated. With this expedient one recovers exactly the relaxed OT formulation in Eq. (S4). In fact, all the transport paths that not not flow into one of the n nodes of Image 2 penalize the cost by traversing the edges connected to u_2. From conservation of mass one can see that these transport paths satisfy $P_{ju_2}^a = H_{ja} - (1/n)\sum_i G_{ia}$, $\forall j \in V_2$. Thus, summing over a and j returns exactly $\sum_{aj} P_{ju_2}^a = ||\sum_j H_j - \sum_i G_i||_1$, with the 1-norm taken over the commodities. This is precisely the penalization we introduced in Eq. (S4).
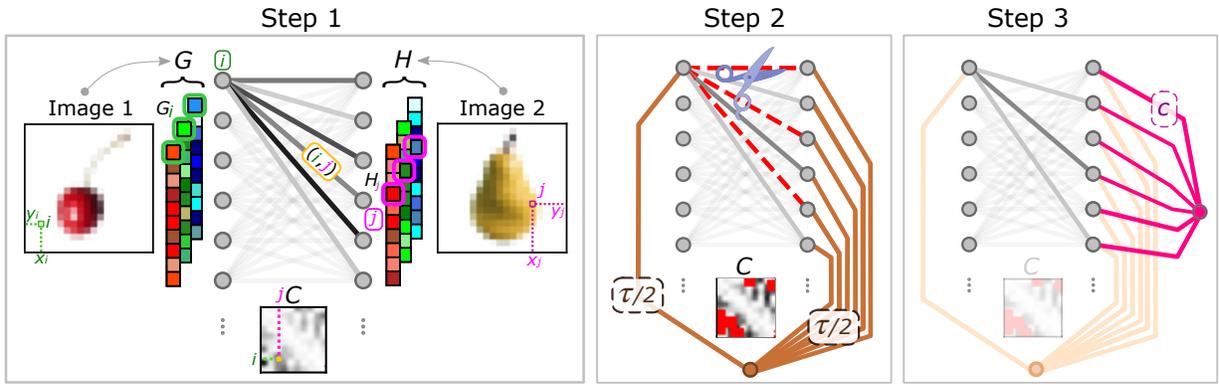
FIG. S1. Detailed construction of transport networks. Step 1: conversion of colored images to tensors and construction of the first complete bipartite graph. Step 2: trimming of expensive edges and addition of the transshipment node, u_1, with its links (in brown). Step 3: relaxation of mass balance with the addition of the second auxiliary node, u_2, together with its links (in magenta).

EQUIVALENCE BETWEEN MULTICOMMODITY DYNAMICS AND OT SETUP

With the following derivations (similar to [28, 29]), we show that asymptotic solutions of Eqs. (2)-(3) are equivalent to minimizers of Eq. (1). This implies that by solving the multicommodity dynamics we find a solution of the multicommodity OT minimization problem. More practically, for a given pair of images, running a numerical scheme on Eqs. (2)-(3) allows us to compute $\lim_{t\to\infty} P(t) = P^\star$, hence $J_\Gamma^\star = J_\Gamma|_{P=P^\star}$, and use the latter as a measure of similarity between them.

More in detail, we first demonstrate the equivalence between stationary solutions of the multicommodity dynamics and minimizers of the multicommodity OT problem introducing a second accessory minimization problem. Stationary solutions are proven to be asymptotes of Eq. (3) only afterwards, with the introduction of a Lyapunov functional for the multicommodity dynamics.

Stationary solutions of the multicomodity dynamics and OT minimizers

Initially, we observe that stationary solutions of the multicommodity dynamics satisfy the relation

$$x_e = ||P_e||_2^{2/(1+\gamma)} \quad \forall e \in E, \tag{S5}$$

that one can derive setting the left hand side of Eq. (3) to zero, defining $P_e^a = x_e(\phi_i^a - \phi_j^a)/C_e$ for $e = (i, j)$, and $\gamma = 2 - \beta$. We recover an scaling identical to Eq. (S5) introducing the following auxiliary constrained minimization problem:

$$\min_{x,P} \left\{ \frac{1}{2} \sum_e \frac{C_e}{x_e} ||P_e||_2^2 + \frac{1}{2\gamma} \sum_e C_e x_e^{\gamma} \right\} \tag{S6}$$

$$\text{s.t.} \quad \sum_e B_{ie} P_e^a = S_i^a \quad \forall i \in V, a = 1, \dots, M. \tag{S7}$$

In fact, differentiating with respect to x_e the objective function in Eq. (S6), and setting the derivatives to zero, yields

$$-\frac{C_e}{x_e^2} ||P_e||_2^2 + C_e x_e^{\gamma-1} \overset{!}{=} 0 \quad \longrightarrow \quad x_e = ||P_e||_2^{2/(1+\gamma)} \quad \forall e \in E. \tag{S8}$$

Noticeably, Eq. (S6) admits a straightforward physical interpretation. In fact, the first term $J = (1/2) \sum_e C_e ||P_e||_2^2 / x_e$ is Joule's first law. Particularly, transport paths can be thought of as fluxes of mass transported through the edges of a capacitated network with resistances $r_e = C_e/x_e$. While the second term, $W_\gamma = (1/2\gamma) \sum_e C_e x_e^{\gamma}$, is the cost needed to build the network infrastructure. The constraints in Eq. (S7)—identical to Eq. (2)—are equivalent to Kirchhoff's law, enforcing conservation of mass.

Most remarkably, the scaling of Eq. (S8) can be also recast in Eq. (S6) to find that $J_\Gamma = J + W_\gamma$ (neglecting multiplicative constants). This connects the multicommodity dynamics with the objective function of Eq. (1). In detail,

$$J + W_\gamma = \frac{1}{2} \sum_e \frac{C_e}{x_e} ||P_e||_2^2 + \frac{1}{2\gamma} \sum_e C_e x_e^{\gamma} \tag{S9}$$

$$\overset{\text{Eq. (S8)}}{=} \frac{1}{2} \sum_e C_e ||P_e||_2^{2\gamma/(1+\gamma)} + \frac{1}{2\gamma} \sum_e C_e ||P_e||_2^{2\gamma/(1+\gamma)} \tag{S10}$$

$$\overset{\Gamma = 2\gamma/(1+\gamma)}{=} \frac{1}{\Gamma} \sum_e C_e ||P_e||_2^{\Gamma} \tag{S11}$$

$$= \frac{1}{\Gamma} J_\Gamma(G, H). \tag{S12}$$

To complete the mapping between the multicommodity dynamics and the minimization setup, we show that the space of transport tensors $\Pi(G, H)$ is exactly the same space defined by Eq. (S7). This can be seen with the following chain of equalities:

$$\sum_k P_{ik}^a - \sum_j P_{ji}^a = G_i^a - H_i^a \quad \forall i \in V, a = 1, \dots, M \tag{S13}$$

$$\sum_k P_{e=(i,k)}^a - \sum_j P_{e=(j,i)}^a = S_i^a \qquad \forall i \in V, a = 1, \dots, M \tag{S14}$$

$$\sum_e B_{ie} P_e^a = S_i^a \qquad \forall i \in V, a = 1, \dots, M. \tag{S15}$$

Here we take the difference between the OT constraints of $\Pi(G, H)$ in Eq. (S13), we then use the definition of S in Eq. (S14), and compact the plus and minus signs using the signed incidence matrix B in Eq. (S15). This allows us to recover Kirchhoff's law as formulated in Eq. (S7) and Eq. (2).

Multicommodity dynamics asymptotes: Lyapunov functional

We complete our discussion introducing the Lyapunov functional for Eq. (3) proposed in [28, 29]. The functional reads:

$$\mathcal{L}_\gamma[x] = \frac{1}{2} \sum_{ai} \phi_i^a[x] S_i^a + \frac{1}{2\gamma} \sum_e C_e x_e^{\gamma}, \tag{S16}$$

and it is a multicommodity generalization of that originally introduced in [36]. This is a well-defined Lyapunov functional for the multicommodity dynamics, in fact, along a curve $x(t)$ solution of Eq. (3),

$$\frac{d\mathcal{L}_\gamma[x(t)]}{dt} \leq 0. \tag{S17}$$

With the equality satisfied if and only if $x(t)$ is a stationary point of Eq. (3). This can be shown as follows. We claim that

$$\frac{\partial \mathcal{L}_\gamma}{\partial x_e} = \frac{C_e}{2}\left(x_e^{\gamma-1} - \frac{||\phi_i - \phi_j||_2^2}{C_e^2}\right) \quad \forall e = (i,j) \in E. \tag{S18}$$

This equality can be retrieved differentiating both sides of Eq. (2) by x_e, thus obtaining

$$\sum_j \frac{\partial L_{ij}}{\partial x_e}\phi_j^a + \sum_j L_{ij}\frac{\partial \phi_j^a}{\partial x_e} = 0 \quad \forall i \in V, e \in E, a = 1, \ldots, M, \tag{S19}$$

$$\sum_j L_{ij}\frac{\partial \phi_j^a}{\partial x_e} = -\sum_j B_{je}(1/C_e)B_{ie}\phi_j^a \quad \forall i \in V, e \in E, a = 1, \ldots, M. \tag{S20}$$

Then, multiplying Eq. (S20) by ϕ_i^a and summing over i one gets

$$\sum_{ij} \phi_i^a L_{ij}\frac{\partial \phi_j^a}{\partial x_e} = -\sum_{ij} \phi_i^a B_{ie}(1/C_e)B_{je}\phi_j^a \quad \forall e \in E, a = 1, \ldots, M, \tag{S21}$$

further summing over a yields

$$\frac{\partial}{\partial x_e}\left(\sum_{aj} S_j^a \phi_j^a\right) = -C_e\frac{||\phi_i - \phi_j||_2^2}{C_e^2} \quad \forall e = (i,j) \in E, \tag{S22}$$

where in the left hand side of Eq. (S22) we used Eq. (2). From Eq. (S22) the equality in Eq. (S18) follows immediately. Now, thanks to Eq. (S18) we can prove that the Lie derivative of the functional is less than or equal to zero. In fact,

$$\frac{d\mathcal{L}_\gamma}{dt} = \sum_e \frac{\partial \mathcal{L}_\gamma}{\partial x_e}\frac{dx_e}{dt} \tag{S23}$$

$$\overset{\text{Eq. (S18)}}{=} \sum_e \frac{C_e}{2}\left(x_e^{\gamma-1} - \frac{||\phi_i - \phi_j||_2^2}{C_e^2}\right)\frac{dx_e}{dt} \tag{S24}$$

$$\overset{\text{Eq. (3),}\gamma=2-\beta}{=} -\sum_e \frac{C_e}{2}x_e^{2-\gamma}\left(x_e^{\gamma-1} - \frac{||\phi_i - \phi_j||_2^2}{C_e^2}\right)^2 \leq 0. \tag{S25}$$

With the equality in Eq. (S25) that is recovered if and only if (i) $x_e = 0$, or (ii) the scaling in Eq. (S8) holds.

Finally, we show that the Lyapunov is identical to the total sum of dissipation and transport cost, i.e., $\mathcal{L}_\gamma = J + W_\gamma$. This can be done multiplying both sides of Eq. (2) by ϕ_i^a and then summing over i and a, namely

$$\sum_{aiej} \phi_i^a B_{ie}(x_e/C_e)B_{je}\phi_j^a = \sum_{ai} \phi_i^a S_i^a \tag{S26}$$

$$\sum_e \frac{C_e}{x_e}||P_e||_2^2 = \sum_{ai} \phi_i^a S_i^a \tag{S27}$$

where we used $P_e^a = x_e(\phi_i^a - \phi_j^a)/C_e$, for $e = (i,j)$. This allows us to conclude.

In summary, we showed that the multicommodity dynamics admits a well-defined Lyapunov functional, which is equivalent to the sum of a dissipation and an infrastructure cost. These two contributions, which are jointly minimized by Eq. (3), when evaluated along their minimizers correspond to the multicommodity OT cost J_Γ of Eq. (1). Introducing the Lyapunov functional is crucial to formally show that asymptotics of the dynamics are equivalent to minimizers of the cost, namely $\lim_{t\to\infty} P(t) = P^\star$.

Lastly, we remark the effect of γ (resp. β) on the minimization problem. In the setting where $\gamma > 1$ ($\beta < 1$) the functional $\mathcal{L}_\gamma$ is convex, with one unique minimizer. For $\gamma < 1$ ($\beta > 1$) the functional landscape becomes rugged and strongly non-convex, with multiple minimizers each correspondent to a local minima of the cost. Hence, in this second scenario, running Eq. (3) permits to converge in a stationary point, which however may not be its global minimum.

CROSS-VALIDATION: FLOWERS DATASET

We perform a 4-fold cross validation on both parameters used for the construction of the ground cost, θ and τ, and on algorithms' regularization parameters, β and ε. We briefly summarize it in this section.

The JF30 Dataset [32] is made of 1,479 elements, divided in 30 classes. First, we separate it into two subsets: *train* and *test*, with classes' frequencies being the same in these subsets as in the entire dataset. To cross-validate our methods, we further separate the train set into 4 folds of equal size, each to be used in turn as a validation set. More in detail, each experiment is executed fixing the validation fold and an image belonging to it, then, the Optimal Transport costs $J_\Gamma^\star$ between such image all the other images in the train set—made of the other three folds—is calculated. This procedure is repeated for all images in the validation set, and swapping each of the 4 train folds as validation set. We use a k-nearest neighbors classifier over $J_\Gamma^\star$ to assign to an image in the validation set its label, that is, for each validation image we consider the k train samples with lowest $J_\Gamma^\star$, and label the validation sample with the most frequent class among these k. This allows us to calculate the classification accuracy of a given fold, and then to average the accuracy over the 4 permutations of the validation and train set. The total amount of experiments we ran in order to cross-validate the model is approximately 50,000.

Results are shown in Fig. S2 and Fig. S3. These depict the average accuracy of: (a) the multicommodity ($M = 3$) dynamics; (b) the unicommodity ($M = 1$) dynamics, both for $\beta \in \{0.5, 0.75, 1, 1.25, 1.5\}$; (c) Sinkhorn algorithm on colored images (Sinkhorn RGB); and (d) Sinkhorn algorithm on grayscale images (Sinkhorn GS), for $\varepsilon \in \{100, 250, 500, 750, 1000, 2500\}$. Letters in parentheses refer to those of Fig. S2 and Fig. S3. The regularization parameters are validated together with $\tau \in \{0.1, 0.125\}$ and $\theta \in \{0, 0.25, 0.5, 0.75\}$. Both multicommodity and unicommodity dynamics are have initial conditions $x_e(0) = 1, \forall e \in E$.

All figures displayed in Fig. S2 and Fig. S3 correspond to highest accuracies returned by the k-NN classifier, with $k = 1, 2, \ldots, 20$. Observing the results, one can see that best performances are attained at $(\tau, \theta, \beta) = (0.125, 0.25, 1)$ for the multicommodity dynamics, and at $(\tau, \theta, \beta) = (0.125, 0.25, 1.25)$ for the unicommodity dynamics.

Noticeably, the accuracy monotonically increases (resp. decreases) with β for a fixed value of θ, namely $\theta = 0$ (resp. $\theta = 0.25$). This can be addressed to the fact that, when no color information is taken into account in the construction of the ground metric ($\theta = 0$), it is more advantageous to consolidate transport paths on cheap edges correspondent to pixels whose positions are close, thus choosing a larger β. On the other hand, introducing colors in C ($\theta > 0$), and thus creating a more disordered ground cost matrix, favors distributing transport paths on the network (See Model Interpretability Section in this SM).

Remarkably, τ also has an impact on the classification accuracy of our algorithms: the larger we set its value to be—thus trimming less edges from the transport network—the more accurate the classification becomes. This behavior is evidently different for Sinkhorn algorithm, as explained here below.
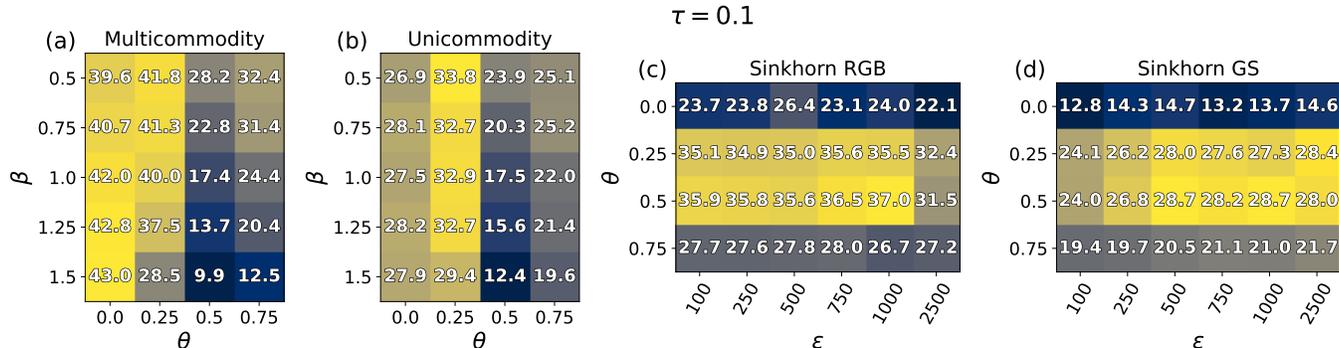
FIG. S2. Cross-validation results for $\tau = 0.1$. Figures are accuracy values obtained with the 4-fold cross validation on JF30. Cells are colored with a darkest-to-brightest palette based on the accuracies. Subplots correspond to: (a) the multicommodity dynamics, (b) the unicommodity dynamics, (c) Sinkhorn on colored images, and (d) Sinkhorn on grayscale images.

Cross-validation of Sinkhorn algorithm is taken a step further. Motivated by the classification accuracy drop observed in Fig. S2, Fig. S3 [(c), (d)] when enlarging the trimming threshold from $\tau = 0.1$ to $\tau = 0.125$, we fix θ and ε to the best values in Fig. S2 [(c), (d)], and progressively reduce τ. Results are shown in Fig. S4 (a) for Sinkhorn on grayscale images, and in Fig. S4 (b) for Sinkhorn on colored images.

Notice that both Sinkhorn GS and Sinkhorn RGB returns bell-shaped curves when changing τ. In particular,
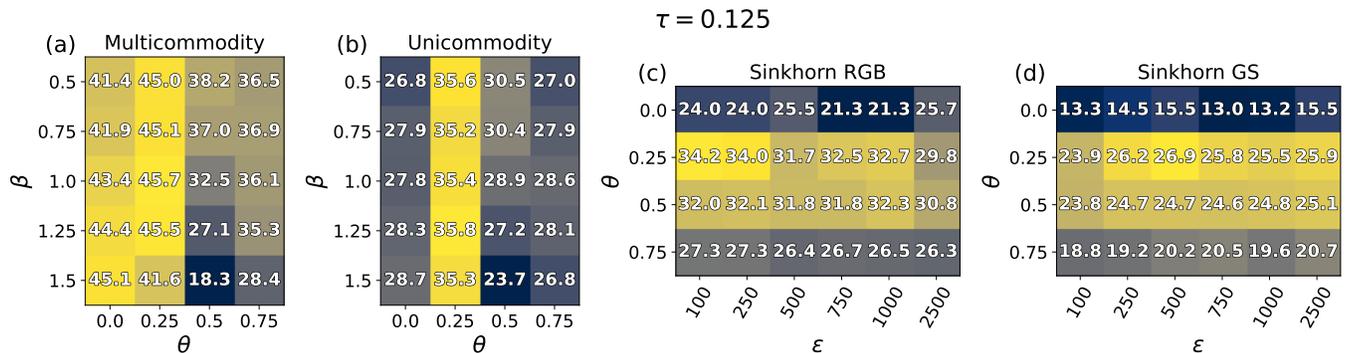
$\tau = 0.125$

FIG. S3. Cross-validation results for $\tau = 0.125$. Figures are accuracy values obtained with the 4-fold cross validation on JF30. Cells are colored with a darkest-to-brightest palette based on the accuracies. Subplots correspond to: (a) the multicommodity dynamics, (b) the unicommodity dynamics, (c) Sinkhorn on colored images, and (d) Sinkhorn on grayscale images.

low classification accuracy is attained when strongly reducing τ, as well as when the trimming threshold is high (approximately $\tau \geq 0.5$). In the first case, many elements of the ground cost matrix are cut, and not enough information is taken into account into the OT setup to properly perform classification. In the second, too much noise is included in into C, which also negatively affects classification. More in detail, we observe in Fig. S4 (a, inset), that Sinkhorn GS performs best when $\tau = 0.05$. For Sinkhorn RGB, i.e. Fig. S4 (b, inset), there is a plateau for all values of the threshold within the interval $[0.05, 0.1]$.

These observations lead us to the choice of $\tau = 0.05$ for Sinkhorn GS, that we re-cross-validate ranging $\theta \in \{0.25, 0.5\}$ and $\varepsilon \in \{100, 250, 500, 750, 1000, 2500\}$. Looking at the results in Fig. S5, we note that optimal parameters for Sinkhorn GS are $(\theta, \varepsilon) = (0.25, 500)$ and $(\theta, \varepsilon) = (0.5, 500)$, which return identical classification accuracy.

As for Sinkhorn RGB, we fix the trimming threshold at the two ends of the plateau in Fig. S4 (b, inset), $\tau = 0.05$ and $\tau = 0.1$, and re-cross-validate them with $(\theta, \varepsilon) = (0.25, 100)$ and $(\theta, \varepsilon) = (0.5, 1000)$. Here, we choose two disparate values of θ and ε not being able to observe a clear relation between these two variables in Fig. S2 (c) and Fig. S3 (c). Namely, $\varepsilon = 100$ (low) and $\theta = 0.25$ perform better for $\tau = 0.125$, in contrast to $\varepsilon = 1000$ (high) and $\theta = 0.5$ for $\tau = 0.1$. Results are in Table S1, optimal parameters are $(\theta, \tau, \varepsilon) = (0.25, 0.05, 100)$.
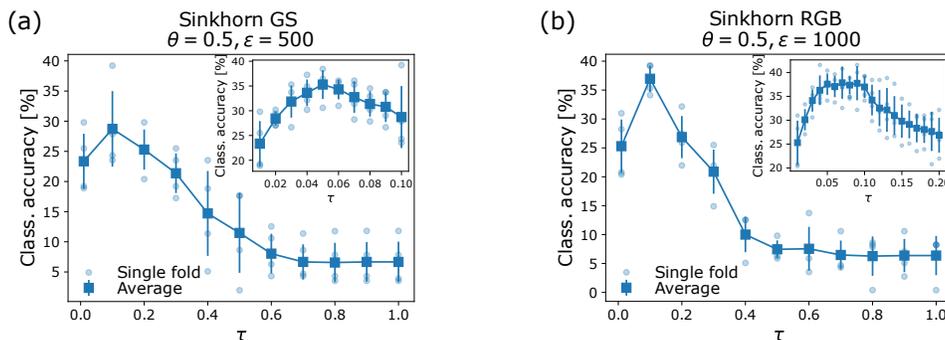
FIG. S4. Sinkohrn's cross-validation results varying τ. Subplots correspond to: (a) Sinkhorn GS, (b) Sinhorn RGB. In each subplot, circular markers correspond to the accuracy values of each fold, instead squares and bars represent their average and standard deviations. In the insets, we refined the grid of τ in an interval of interest, where classification accuracy is peaked.

EXPERIMENTAL DETAILS: FRUITS DATASET

Here, we describe in detail the experimental setup designed for the Fruit Dataset (FD) [33]. FD consists of 163 images of 15 fruit types. We split the whole dataset into *train* and *test* sets, each with 70% and 30% of the available images, respectively. As for the other dataset, classes' frequencies are the same in these subsets as in the entire dataset. Given the rather small size of this dataset, we directly perform classification comparing train and test. All the experiments have been executed with the two best performing parameter configurations of ε and θ, cross-validated

$$\tau = 0.05$$

Sinkhorn GS

θ	100	250	500	750	1000	2500
0.25	26.1	31.5	35.3	35.1	35.1	32.2
0.5	28.1	31.7	35.3	35.0	34.9	29.9

ε
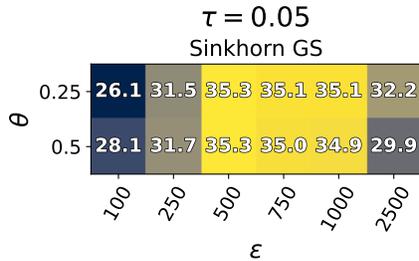
FIG. S5. Refined cross-validation results Sinkhorn GS and $\tau = 0.05$. Figures are accuracy values obtained with the 4-fold cross validation on JF30. Cells are colored with a darkest-to-brightest palette based on the accuracies.

Algorithm	Hyperparameters				Class. accuracy
	θ	τ	ε	k	[%] ($\uparrow$)
Sinkhorn RGB	0.25	0.05	100	1	58.4
	0.5	0.05	1000	1	53.6
	0.25	0.1	100	1	53.2
	0.5	0.1	1000	1	49.0

TABLE S1. Refined cross-validation results for Sinkhorn RGB. Rows are sorted (from bottom to top) using the average percent accuracy obtained with the 4-fold validation (from worst to to best).

on JF30, for Sinkhorn-based methods. These are: $(\theta, \varepsilon) = (0.25, 500)$, $(\theta, \varepsilon) = (0.5, 500)$ for Sinkhorn GS [see Fig. S2 (d)], and $(\theta, \varepsilon) = (0.5, 1000)$, $(\theta, \varepsilon) = (0.5, 750)$ for Sinkhorn RGB [see Fig. S2 (c)]. For our dynamics, we selected the two best performing values of β, for $\theta = 0$ and $\theta = 0.25$. Namely, $(\theta, \beta) = (0.25, 1)$, $(\theta, \beta) = (0.25, 1.5)$ for the multicommodity dynamics [see Fig. S3 (a)], and $(\theta, \beta) = (0.25, 1.25)$, $(\theta, \beta) = (0, 1.5)$ for the unicommodity dynamics [see Fig. S3 (b)]. The trimming threshold is ranged in $\tau \in \{0.04, 0.05, 0.06, 0.07\}$.

IMAGE PREPROCESSING

The elements of both datasets are processed in the following way. First, each image is coarsened with an average pooling, the only input needed for this step is the size of the square mask, `ms`. Its stride is in fact set to `stride = ms`, and the padding to `pad = 0`. All images were conveniently trimmed so that both their widths and heights are divisible by the pooling mask size. We set `ms = 40` for JF30, and `ms = 30` for FD. Furthermore, we smooth the images using a Gaussian filter on each color channel, with standard deviation $\sigma = 0.5$.

Moreover, to convert colored images into grayscale ones, which are given as input to Sinkhorn GS and to our unicommodity dynamics ($M = 1$), we preprocess them as follows. Let (R, G, B), be the three color channels composing each pixel of a colored image, these are converted into a unique channel (its grayscale counterpart), whose intensity I is calculated with the weighted sum $I = 0.2125R + 0.7154G + 0.0721B$. The weights correspond to those used by cathode-ray tube (CRT) phosphors as they are more suitable to represent human perception of red, green and blue than equally valued weights [37].

Color distributions of images

As shown in Table I, for the multicommodity and the unicommodity dynamics, optimal values of the trimming threshold τ are much lower in the Fruit Dataset [33] than in the Jena Flowers 30 Dataset [32]. This can be addressed to the fact that color distributions of fruits, belonging to the first dataset, are drastically light-tailed compared to those of flowers in the second dataset. Thus, the cost C is naturally noisier in the latter case, and a larger trimming is necessary to remove such noise from classification.

Most of the noise in pictures of flowers comes from the background. In fact, while all flowers are photographed in nature, fruits are depicted on a white background. This can be seen in Fig. S6 (a)-(d). In subplot (a) we show four images randomly sampled from the Fruit Dataset, in (c) four random samples of the Jena Flowers 30 Dataset. In (b) and (d) we plot the average color intensity of the RGB channels for 100 random samples belonging to the two

datasets. Here, the histograms in (b) are relative to the fruits, those in (d) to the flowers. From the plots it can be clearly seen that the color distributions of Fig. S6 (b) are starkly peaked around $(R, G, B) = (255, 255, 255) =$ white in standard RGB encoding.
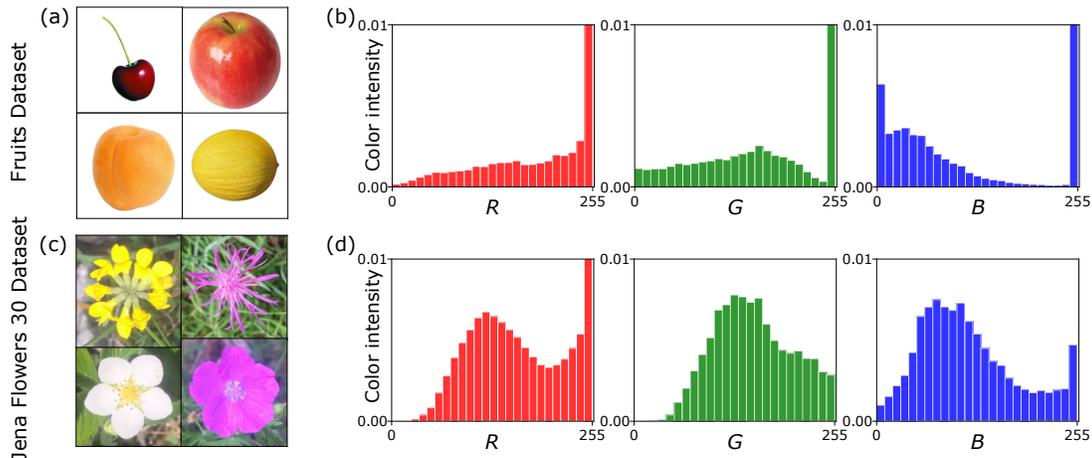
FIG. S6. Color distributions in the two datasets. Subplots (a), (d) are relative to FD, subplots (c), (d) to JF30. In (a), (c) we plot four random images drawn from the two datasets. In (b), (d) the average color intensities (properly normalized to sum to one) of 100 random samples extracted from the two datasets. The plots correspond to red $= R$, green $= G$, and blue $= B$.

SINKHORN BENCHMARKS

In our experiments, we compare the multicommodity and unicommodity dynamics against Sinkhorn algorithm, popularized by the seminal work of Cuturi [16]. The idea of Sinkhorn is to regularize the standard OT problem by adding an entropic barrier to the cost function. More in detail, and following the notation adopted in our manuscript, the minimization problem proposed in [16] is:

$$\min_{\substack{P \text{ s.t. } \sum_j P_{ij} = g_i \\ \sum_i P_{ij} = h_j}} \left\{ \sum_{ij} P_{ij} C_{ij} - \varepsilon h(P) \right\}, \qquad h(P) = -\sum_{ij} P_{ij} \log P_{ij}. \tag{S28}$$

Here transport paths P, which generally lie in the polyhedral set described by the constraints $\sum_j P_{ij} = g_i \, \forall i$ and $\sum_i P_{ij} = h_j \, \forall j$, are smoothed by the entropy $h(P)$. This trick makes the optimization problem strictly convex, and permits to solve it with a very efficient matrix scaling algorithm—Sinkhorn's fixed point iteration.

We generalize the problem in Eq. (S28) in order to take in account transport tensors, G and H, which carry information of multiple color channels, and transport paths P. In detail, we propose the following minimization problem for each commodity—color channel—a,

$$\min_{\substack{P^a \text{ s.t. } \sum_j P_{ij}^a = G_i^a \\ \sum_i P_{ij}^a = H_j^a}} \left\{ J_{\text{sink}}^a = \sum_{ij} P_{ij}^a C_{ij} - \varepsilon h(P^a) \right\}, \qquad h(P^a) = -\sum_{ij} P_{ij}^a \log P_{ij}^a. \tag{S29}$$

This allows to efficiently compute, using Sinkhorn's scaling, an Optimal Transport path P_{opt}^a for each commodity, together with its correspondent optimal cost $J_{\text{sink,opt}}^a = J_{\text{sink}}^a|_{P^a = P_{\text{opt}}^a}$. Finally, the Optimal Transport cost for colored images is calculated as $J_{\text{sink,opt}}^{\text{RGB}} = (1/3) \sum_{a=1}^{M=3} J_{\text{sink}}^a$.

MODEL INTERPRETABILITY

In this section we discuss the effect that the parameters θ, τ, and β have one the OT setup.

First, we explain the experiment in Fig. S7. We start by sampling two images of the FD dataset belonging to the same class. These images have identical shape, i.e. width w and height h equal to 20. They are displayed on the leftmost part of Fig. S7. From these two images, we obtain the tensors G and H, that are transported in the OT problem. The first, G, is constructed using all the pixels on the 11th row of Image 1, thus its dimension is $m \times M = 20 \times 3$. The same row of Image 2 is used to build H, also in this case its size is $n \times M = 20 \times 3$.

The two tensors enter in Eq. (1) together with a (20×20)-dimensional cost C, which is built with pixels' locations and color information using Eqs. (S1)-(S3). The scope of this discussion is to refine the intuition on these formulas, and on the effect that θ and τ have on C. In Fig. S7 we plot the ground cost C for the two tensors G and H, for $\theta = \{0, 0.25, 0.5, 0.75\}$ and $\tau = \{0.1, 0.25, 0.5, 1\}$. All entries $C_{ij} > \tau$—which correspond to those edges that are trimmed from the transport network—are colored in white.

Notice that for $\theta = 0$ all costs are symmetric. Indeed in this case $C_{ij} = \min\{Y_{ij}, \tau\}$, with Y that is containing the Euclidean distances between pixels' coordinates, i.e. Eq. (S2). Here, decreasing the trimming threshold τ progressively sparsifies the banded matrices drawn in the first column of Fig. S7, with limit cases being $C = \mathrm{diag}[C_{0,0}, C_{1,1}, \ldots, C_{19,19}]$—for τ sufficiently small, and $C = Y$—for $\tau \geq \max_{ij} Y_{ij}$. On the other hand, the symmetry is gradually broken as θ is increased, namely, when colors of the images are are used to build into C. This is clearly depicted in Fig. S7, where the heatmaps get progressively disordered for larger values of θ (from left to right).
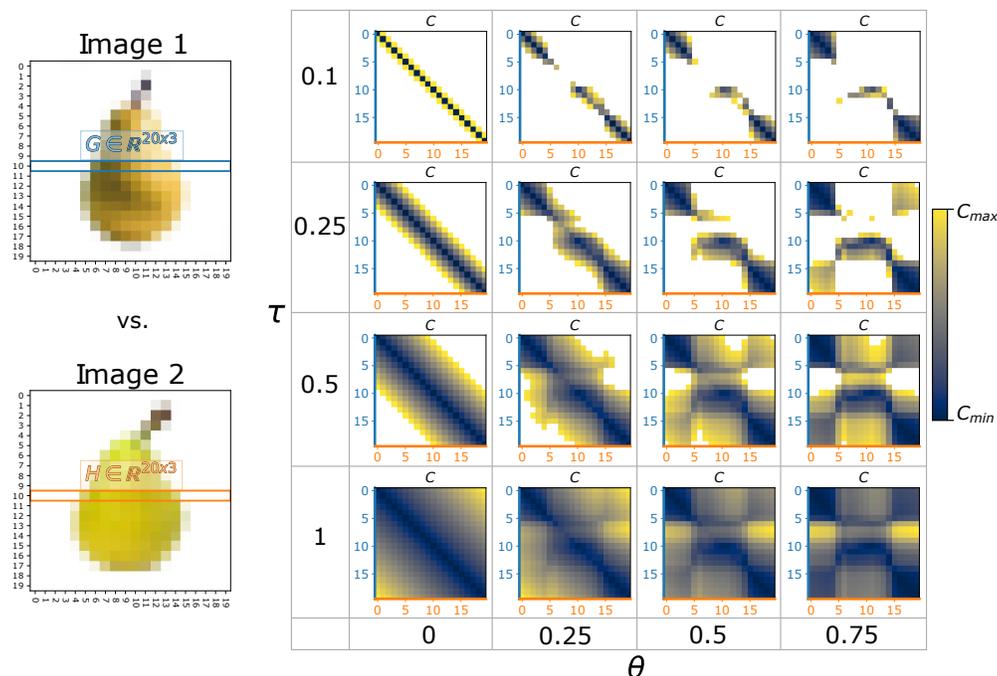
FIG. S7. Effect of θ and τ on OT. On the left, we display the two samples used to build the ground costs C. Highlighted rows in blue and orange are those considered to extract G and H. On the right side of the panel we plot C for $\theta = \{0, 0.25, 0.5, 0.75\}$ and $\tau = \{0.1, 0.25, 0.5, 1\}$. White regions correspond to trimmed values, i.e. entries of C that are larger than τ.

To further expand this discussion, we design a second experiment, schematically represented in Fig. S8. Here, we solve the OT problem between two tensors, G and H, built similarly to those of Fig. S7. Particularly, we consider three central pixels of the 11th rows of Image 1 and Image 2, as drawn in the leftmost part of the Figure, so that both g^a and h^a are 3-dimensional arrays for all $a = 1, \ldots, 3$, and C is a (3×3)-dimensional matrix.

Depending on the values of θ, the ground cost C is either symmetric ($\theta = 0$), and computed only using pixels' coordinates, or strongly irregular ($\theta = 0.75$), since colors of images are taken into account. In the first case, the transport network connecting the images has also a symmetric structure. Here, elements along the diagonal of the cost—correspondent to horizontal edges connecting orange and blue nodes with the same index—are much cheaper than all the other entries. This is due to the fact that the Euclidean distance between two pixels with the same position is zero (practically set to a safety default value $\epsilon = 10^{-5}$). Conversely, in the second case, introducing colors in the ground cost translates into having higher values along the diagonal elements of C. Here, colors—which distribute more smoothly on images—smooth out the cost as well, whose entries are more homogeneous.

As shown in Fig. S2 and Fig. S3, taking a purely Euclidean ground cost C, i.e., $\theta = 0$, returns higher classification accuracy when $\beta = 1.5$. Instead, building C mostly with color information, thus setting $\theta = 0.75$, favors $\beta = 0.5$. We address this tendency to the effect that β has on transport paths' consolidation, and we represent it on the rightmost portion of Fig. S8, where we plot the Optimal Transport paths $\{P^1, P^2, P^3\}$ obtained running Eqs. (2)-(3) on the OT setup just discussed. In detail, for $\theta = 0$, horizontal edges in the transport network are much cheaper than the others, therefore strong consolidation of transport paths ($\beta = 1.5$) benefits classification. Conversely, since for $\theta = 0.75$ the entries of C are more homogeneous, distributing transport paths ($\beta = 0.75$) naturally reflects the topology of the transport network and allows to achieve better classification performances.

Lastly, we remark that transport paths do not pass through any transshipment edge (colored in brown in Fig. S8) since τ is conveniently set to be sufficiently large. The auxiliary edges for Kirchhoff's law relaxation (colored in magenta) are instead traversed by transport paths since G and H are not normalized.
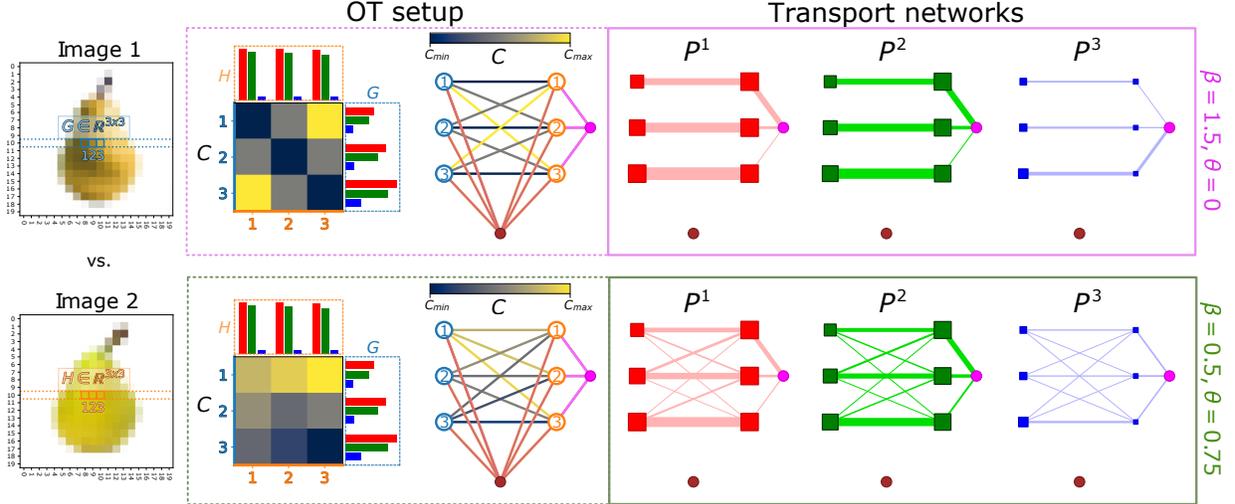
FIG. S8. Effect of β on OT. In the leftmost portion of the panel we plot Image 1 and Image 2, used in the OT problem. From these we extract the (3×3)-dimensional tensors G and H. These are drawn together with a heatmap of the cost C, and with the correspondent transport network. Color scales of edges and of entries of C are identical. We also use the same numbering and color scheme for tensors' entries, indexes of C, and network nodes. Brown and magenta auxiliary nodes and edges are added after trimming, and after relaxing Kirchhoff's law. On the right side of the panel we plot the transport network again, but with edge thickness proportional to the Optimal Transport paths retrieved from Eqs. (2)-(3), and with colors correspondent to those of the commodities a. Node sizes are proportional to the values of g^a and h^a, for $a = 1, \ldots, 3$.

COMPUTATIONAL COST

Analytical discussion

Considerable effort has been spent to reduce the high complexity burden of OT problems. The $O(|V|^2/\varepsilon^3)$ baseline of Sinkhorn algorithm [16, 38, 39], where $|V|$ is the size of the histograms transported and ε the parameter enforcing entropic regularization, is constantly improved. Notable recent results are the class of stochastic optimization algorithms proposed in [40], that have been ameliorated using greedy alternatives [41] to achieve ε-approximation of the 1-Wasserstein distance between two probability distributions in $O(|V|^2/\varepsilon^2)$ arithmetic operations [22]. Recently, an Adaptive Primal-Dual Accelerated Gradient Descent (APDAMD) scheme with complexity $O(\min\{|V|^{9/4}/\varepsilon, |V|^2/\varepsilon^2\})$ for the same ε-perturbed problem has been presented in [23].

In principle, our multicommodity method has a computational complexity of order $O(M|V|^2)$ for complete transport graph topologies, i.e., when edges in the transport network K are assigned to all pixels' pairs. Nonetheless, we achieve a substantial decrease in complexity by sparsifying the graph with the trimming procedure of [19, 20]. Similarly to [19], the final complexity of our algorithm is $O(M|V|)$. This improvement can be formally justified as follows, we start from a complete bipartite graph with $|E| = |V|^2/4$ (for simplicity $m = n = |V|/2$ is assumed). First, we trim expensive links, and reduce the number of edges of the transport network to $\langle K \rangle |V| + |V|$, where $\langle K \rangle$ is the average number of edges connected to a node that are not trimmed by τ, and the second term $|V|$ counts the number of inflowing

and outflowing transshipment links. Second, we add $|V|/2$ links to the transport network to enforce Kirchhoff's law penalization, so that the final number of links amounts to $|E| = |V|(\langle K \rangle + 3/2)$, which is linear with respect to the number of nodes.

Empirically, we observe that running Eqs. (2)-(3), most of the entries of x decay to zero after a few steps, producing a progressively sparser weighted Laplacian $L[x]$. This permits a faster computation of the Moore-Penrose inverse $L^\dagger[x]$, and of the least-square potentials $\phi_i^a = \sum_j L_{ij}^\dagger[x] S_j^a$ thereon. An experimental thorough analysis of convergence properties of the OT dynamics has been done in [42]. Here, it is shown that for $\beta = 1$, the Optimal Transport paths of the unicommodity OT problem on sparse topologies amount to the solution of z linear systems as in Eq. (2), with $O(1) < z < O(|E|^{0.36})$. This bound has been found using a backward Euler scheme combined with the inexact Newton-Raphson method for the update of x, and solving Kirchhoff's law using an algebraic multigrid method.

Experimental runtimes benchmarking against Sinkhorn

We compare the runtime performances of the multicommodity dynamics of Eqs. (2)-(3) against the regularized Sinkhorn algorithm of [43, 44], implemented in POT: Python Optimal Transport [31], and for which we set the convergence threshold to $\widetilde{\varepsilon}_{\text{sink}} = 0.01$. Our implementation uses a forward Euler scheme for the discretization of Eq. (3), and a sparse direct linear solver (UMFPACK) for Eq. (2). Our code was run until convergence, achieved if $(J_\Gamma(n+1) - J_\Gamma(n))/\Delta t < \widetilde{\varepsilon}_{\text{dyn}}$, i.e. when the relative cost difference evaluated at two consecutive iteration is below $\widetilde{\varepsilon}_{\text{dyn}} = 1$. We set the discretization time step $\Delta t = 0.5$.

All codes are executed on 20 pairs of images, randomly sampled from the Jena Flowers 30 Dataset [32] and the Fruit Dataset [33]. We compare our multicommodity dynamics ($M = 3$) against Sinkhorn algorithm on colored images, and the unicommodity dynamics ($M = 1$) against Sinkhorn on grayscale images.

Results are shown in Fig. S9. Here, we plot elapsed times for the experiments on JF30 and on FD in the panels (a)-(c) and (d)-(f), respectively. Subplots from left to right represent runtimes for the algorithms executed on grayscale images [(a), (d)], on colored images [(b), (e)], and for Sinkhorn executed in both setups [(c), (f)].

Observing Fig. S9 [(a), (d)], we notice that runtimes for the multicommodity dynamics are larger than Sinkhorn's. Our algorithm converges faster if $\beta < 1$, i.e., when the multicommodity transport cost is convex. Setting $\beta > 1$ negatively affects convergence times. In general, for all values of β, increasing the trimming threshold τ, and thus the average number of edges in the transport networks, leads to slower convergence. Sinkhorn algorithm is not as dependent on $|E|$, e.g., in Fig. S9 (a), runtimes are approximately constant. Moreover, coherently to what expected [16], increasing the effect of the entropic barrier—enlarging ε—makes the algorithm faster. In Fig. S9 [(b), (e)] we observe a similar trend as in Fig. S9 [(a), (d)]. However, in this case Sinkorn algorithm with low regularization, $\varepsilon = 100$, has runtimes comparable to those of our method. Lastly, in Fig. S9 [(c), (f)], we explicitly plot runtimes for Sinkhorn on both colored and grayscale images, for different values of the regularization parameter ε. In general, the algorithm on colored images is slower, and increasing the trimming threshold leads to higher runtimes. Moreover, we observe again that larger value of ε makes the algorithms faster.
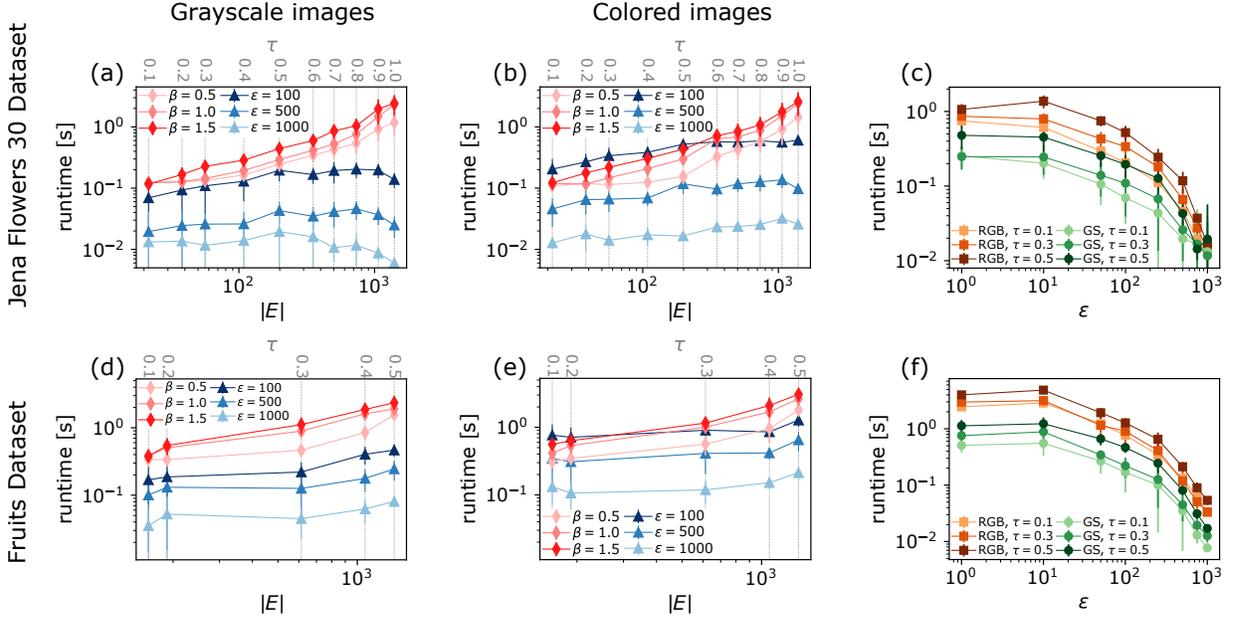
FIG. S9. Runtimes of algorithms. Subplots (a)-(c) are experiments on JF30, subplots (d)-(f) are those on FD. In (a), (b), (c), and (d) we plot with red diamonds runtimes of our dynamics, with $M = 1$ in (a), (d) and $M = 3$ in (b), (e). Blue triangles are denote runtimes of Sinkhorn. Color shades correspond to different values of the regularization paramters. In (c) and (f) we show runtimes of Sinkhorn against ε, with orange and green markers used for colored and grayscale images, respectively. Color shades here denote different values of the trimming threshold τ. Errorbars are standard deviations obtained over 20 random image pairs.