Unbalanced Optimal Transport: A Unified Framework for Object Detection Supplementary Material

A. Optimal Transport Discussion

The *Otpimal Transport* formulation presented throughout the paper is formulated in discrete space. In this section, we present the more general formulation of which the discrete one is a particular case of. We also discuss the Wasserstein distance in the particular context of object detection and the effect of regularization on the uniqueness of the solutions. Only the case of the original OT formulation—or "balanced" case—is covered here.

A.1. Continuous Formulation

More generally, we define Optimal Transport in its continuous form.

Definition 5 (Continuous Optimal Transport)**.** *Given two distributions $\alpha \in \mathscr{P}_+(X)$ and $\beta \in \mathscr{P}_+(Y)$ of same mass $\int \alpha \, \mathrm{d}x = \int \beta \, \mathrm{d}y$, and given an underlying cost function $c : X \times Y \to [0, +\infty]$, we define* Continuous Optimal Transport *as the minimization of a transport cost*

$$\inf \left\{ \int_{X \times Y} c \, \mathrm{d}\gamma : \gamma \in U(\alpha, \beta) \right\}, \tag{8}$$

with admissible solutions, here called transport plans

$$U(\alpha, \beta) = \left\{ \gamma \in \mathscr{P}_+(X \times Y) : \int_Y \mathrm{d}\gamma = \alpha \quad \textit{and} \quad \int_X \mathrm{d}\gamma = \beta \right\}. \tag{9}$$

If a minimum exists, it is called the optimal transport plan $\hat{\gamma}$.

We replace the probability simplex Δ^N by the space on probability distributions $\mathscr{P}_+(X)$ on X. The transport plans are the set of joint probability distribution $\gamma \in \mathscr{P}_+(X \times Y)$, whose marginal distributions are α and β. The discrete formulation (Definition 1) is a particular case where $\alpha = \sum_i \alpha_i \delta_{\hat{\boldsymbol{y}}_i}$, $\beta = \sum_j \beta_j \delta_{\boldsymbol{y}_j}$ and the cost $c = \mathcal{L}_{\text{match}}$. In this case, a minimum always exists.

A.2. Wasserstein Distance

This infimum defines a distance between α and β, called the Wasserstein distance $\mathcal{W}_p(\alpha, \beta)$, provided that the underlying cost function is also a distance $c = d^p$ up to some exponent $p \in [1, +\infty[$. In our case, $\mathcal{L}_{\text{match}}$ is not a distance. More formally, sum of distances are distances. The ℓ^1 norm is a distance, and $1 - \text{IoU}$, or $1 - \text{GIoU}$ also are [7]. However, the cross entropy or the focal loss do not satisfy the triangular inequality or the symmetry properties. In consequence, we cannot talk about a Wasserstein distance here.

Furthermore, interpreting a Wasserstein distance $\mathcal{W}_p(\alpha, \beta)$ would not make much sense even if the underlying matching cost was to be a distance. Indeed, the distributions α and β would be the same at every iteration in our framework. In other words, the distance would always be computed between the same points, but the underlying cost would change and it would be different for each image. Each iteration would be computing the distance of two same points in a changing geometry and each image would have its own evolving geometry.

For completeness, we must mention that the regularized version does not define a distance as $\mathcal{W}_{p,\text{reg.}}(\boldsymbol{\alpha}, \boldsymbol{\alpha}) = -\epsilon \, \mathrm{H}(\boldsymbol{I}_{N_p,N_p}/N_p) > 0$ with $\boldsymbol{I}_{N_p,N_p}$ the identity matrix of size N_p (we refer to [5, 4, 3] for a broader discussion on the subject).

A.3. Uniqueness

We consider here the discrete formulation used throughout the paper. By classical linear programming theory, the non-regularized problem admits a non-unique solution if and only if multiple extreme points minimize the problem. In that case, the set of minimizers is all the linear interpolations between those extreme points. The regularization term however is ϵ-strongly convex; the regularized problem thus always has a unique solution [6].

B. Proofs of the Propositions

In this section, we provide the proofs of Propositions 1 and 2 and enrich them with some insight through a few additional results.

B.1. Hungarian Algorithm

Before providing a proof of the particular equivalence between OT and BM, we first consider a more general result.

Lemma 1. *We consider the rational probability simplex* $\Delta_{\mathbb{Q}}^{N} = \{\boldsymbol{u} \in \mathbb{Q}_{\geq 0}^{N} | \sum_i u_i = 1\}$. *Given an OT problem (Definition 1) with underlying distributions* $\boldsymbol{\alpha} \in \Delta_{\mathbb{Q}}^{N}$ *and* $\boldsymbol{\beta} \in \Delta_{\mathbb{Q}}^{M}$. *Each extreme point of* $\mathcal{U}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ *is comprised of elements, which are multiples of the* common measure *of* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\beta}$*:*

$$\boldsymbol{P} \textit{ is an extreme point of } \mathcal{U}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \qquad \Longrightarrow \qquad \boldsymbol{P} \in \mathrm{CM}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \cdot \mathbb{N}_{\geq 0}^{N \times M}, \tag{10}$$

where the common measure *is the greatest rational such that all non-zero elements of both distributions are multiples of it:*

$$\mathrm{CM}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\mathrm{GCD}\left(\mathrm{LCM}\left([\,\boldsymbol{\alpha},\, \boldsymbol{\beta}\,]\right) / \,[\,\boldsymbol{\alpha},\, \boldsymbol{\beta}\,]\right)}{\mathrm{LCM}\left([\,\boldsymbol{\alpha},\, \boldsymbol{\beta}\,]\right)} \in \mathbb{Q}_{>0}, \tag{11}$$

with $\mathrm{GCD} : \mathbb{N}_{>0}^{N} \to \mathbb{N}_{>0}$ *the greatest common divisor and* $\mathrm{LCM} : \mathbb{N}_{>0}^{N} \to \mathbb{N}_{>0}$ *the lowest common multiple.*

The common measure *extends the* GCD *to non-integers. As an example* $\mathrm{CM}([\,{}^{2}\!/_{3},\ {}^{4}\!/_{5}\,]) = {}^{2}\!/_{15}$ *and* $\mathrm{CM}([\,{}^{2}\!/_{3},\ {}^{5}\!/_{6},\ {}^{4}\!/_{7}\,]) = {}^{1}\!/_{42}$.

Proof. In [1], Corollary 8.1.3, an algorithm is given to build the exhaustive list of extreme points. It comprises only minimum and subtraction operations, which leave the common measure unchanged. □

Corollary 1. *Given the underlying distributions as in Proposition 1, the extreme points of* $\mathcal{U}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ *are comprised only of zeros and* $1/N_p$*:*

$$\boldsymbol{P} \textit{ is an extreme point of } \mathcal{U}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \qquad \Longrightarrow \qquad \boldsymbol{P} \in \{0, 1/N_p\}^{N_p \times (N_g + 1)}. \tag{12}$$

This is a direct consequence of Lemma 1 and the mass constraints directly implying that $P_i \leq 1/N_p$ for all i. In this particular case, there is also an equivalence.

Lemma 2. *Given the underlying distributions as in Proposition 1, the extreme points of* $\mathcal{U}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ *are comprised only of zeros and* $1/N_p$*:*

$$\boldsymbol{P} \textit{ is an extreme point of } \mathcal{U}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \qquad \Longleftrightarrow \qquad \boldsymbol{P} \in \{0, 1/N_p\}^{N_p \times (N_g + 1)} \quad \textit{and} \quad \boldsymbol{P} \in \mathcal{U}(\boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{13}$$

Proof. We consider Corollary 1 and add the fact that such a match $\boldsymbol{P} \in \{0, 1/N_p\}^{N_p \times (N_g + 1)}$ only has one element per row (or prediction if we prefer) to satisfy the mass constraints. Therefore, it cannot be any interpolation of two other extreme points. □

We however also give a more direct proof, based essentially on the same arguments.

Proof. We will first show that the elements of the match $\boldsymbol{P}$ corresponding to any extreme point, can only be $1/N_p$ or 0. Therefore we can consider the associated bipartite graph of the problem: each prediction consists in a node i and each ground truth a node j. Each non-zero value entry of $\boldsymbol{P}$ connects nodes i and j with weight $P_{i,j}$. The solution is admissible if and only if the weight of each node i equals α_i and j equals β_j. A transport plan $\boldsymbol{P}$ is an extreme point if and only if the corresponding bipartite graph only consists in trees, or equivalently, it has no cycle (Theorem 8.1.2 of [1]).

Because the mass constraint must all sum up to one for the predictions, we already know that $P_{i,j} \leq 1/N_p$. We will now proceed *ad absurdum* and suppose that there were to be an entry $0 < P_{i,j} < 1/N_p$ connecting a prediction and a ground truth. In order to satisfy the mass constraints, they would both also have to be connected to another prediction and another ground truth. Similarly, these would also have to be connected to at least one prediction and one ground truth, and so on. They would all form a same graph, or be "linked" together in other words. By consequence, each new connection must be done to yet "unlinked" prediction and ground truth to avoid the formation of a cycle. Considering that there are N_p predictions, there would be at the end at least $2N_p$ edges within the graph. This is incompatible with the fact that there cannot be any cycle (Corollary 8.1.3 of [1]). By consequence, the entries of $\boldsymbol{P}$ must be either 0 or $1/N_p$. □

We can now proceed to prove the said proposition.

Proposition 1. *The Hungarian algorithm with N_p predictions and $N_g \leq N_p$ ground truth objects is a particular case of OT with $\boldsymbol{P} \in \mathcal{U}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \subset \mathbb{R}^{N_p \times (N_g+1)}$, consisting of the predictions and the ground truth objects, with the background added $\{\boldsymbol{y}_j\}_{j=1}^{N_g+1} = \{\boldsymbol{y}_j\}_{j=1}^{N_g} \cup (\boldsymbol{y}_{N_g+1} = \varnothing)$. The chosen underlying distributions are*

$$\begin{aligned} \boldsymbol{\alpha} &= \frac{1}{N_p}[\,\underbrace{1,\ 1,\ 1,\ \ldots,\ 1}_{N_p \text{ predictions}}\,], \qquad (14) \\ \boldsymbol{\beta} &= \frac{1}{N_p}[\ \underbrace{1,\ 1,\ \ldots,\ 1}_{N_g \text{ ground truth objects}}\ ,\ \underbrace{(N_p - N_g)}_{\text{background } \varnothing}\,], \qquad (15) \end{aligned}$$

provided the background cost is constant: $\mathcal{L}_{match}(\hat{\boldsymbol{y}}_i, \varnothing) = c_\varnothing$. In particular for $j \in [\![N_g]\!]$, we have $\hat{\sigma}(j) = \{i : P_{i,j} \neq 0\}$, or equivalently $\hat{\sigma}(j) = \{i : P_{i,j} = 1/N_p\}$.

Proof. We will demonstrate that OT with $\boldsymbol{\alpha} = \frac{1}{N_p}[\ 1,\ 1,\ 1,\ \ldots,\ 1\ ]$ and $\boldsymbol{\beta} = \frac{1}{N_p}[\ 1,\ 1,\ \ldots,\ 1,\ (N_p - N_g)\ ]$ and constant background cost necessarily has the BM as minimal solution. We first observe that because of the linear nature of the problem, there is at least one extreme point that minimizes the total cost. By directly applying Lemma 2, there must be exactly one match per prediction and exactly one match for each non-background ground truth to satisfy the mass constraints. The added background ground truth has $N_p - N_g$ matches. This is equivalent to saying that disregarding the background ground truth, we wave $\sigma \in \mathcal{P}_{N_g}([\![N_p]\!])$ with $\hat{\sigma}(j) = \{i : P_{i,j} = 1/N_p\}$. The proof is concluded by observing that the part of the background in the total transport cost is equal to $\frac{1}{N_p}(N_p - N_g)\, c_\varnothing$ and is constant, hence not influencing the minimum. □

B.2. Minimum Matching with Threshold

Proposition 2 (Matching to the closest)**.** *We consider the same objects as Proposition 1. In the limit of $\tau_1 \to \infty$ and $\tau_2 = 0$, Unbalanced OT (Definition 4) without regularization ($\epsilon = 0$) admits as solution each prediction being matched to the closest ground truth object unless that distance is greater than a threshold value $\mathcal{L}_{match}(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_{N_g+1} = \varnothing) = c_\varnothing$. It is then matched to the background $\varnothing$. In particular, we have*

$$\hat{P}_{i,j} = \begin{cases} \frac{1}{N_p} & \text{if } j = \arg\min_{j \in [\![N_g+1]\!]} \{\mathcal{L}_{match}(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j)\}, \\ 0 & \text{otherwise.} \end{cases} \qquad (16)$$

Proof. By taking the limit of $\tau_1 \to +\infty$ and setting $\epsilon, \tau_2 = 0$, the problem becomes

$$\begin{array}{ll} \arg\min & \left\{ \sum_{i,j=1}^{N_p, N_g+1} P_{i,j} \mathcal{L}_{\text{match}}(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_j) \middle| \boldsymbol{P} \in \mathbb{R}_{\geq 0}^{N_p \times (N_g+1)} \right\}, \\ \text{s.t.} & \sum_j P_{i,j} = 1/N_p \qquad \forall i. \end{array} \qquad (17)$$

We can now see that the choice made in each row is independent from the other rows. In other words, each ground truth object can be matched independently of the others. The minimization is then obtained if, for each prediction (or row), all the weight is put on the ground truth object with minimum cost, including the background. This leads to Eq. (16). □

Corollary 2 (Matching to the closest without threshold)**.** *Provided the background cost is more expensive than any other cost $c_\varnothing > \max\left\{\mathcal{L}_{match}(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j) \middle| i \in [\![N_p]\!] \text{ and } j \in [\![N_g]\!]\right\}$, each prediction will always be matched to the closest ground truth.*

In theory, this a much too strong condition, the background cost can just be greater than the minimum cost for each prediction $\mathcal{L}_{\text{match}}(\hat{\boldsymbol{y}}_i, \varnothing) > \min_j \{\mathcal{L}_{\text{match}}(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j)\}$. In practice, however, this does not change much. It suffices to set the background cost high enough and we are assured to get a minimum. One could also imagine a different background cost for each prediction in order to have a more granular threshold.

C. Scaling Algorithms

We present here the two scaling algorithms: Sinkhorn's algorithm for "'balanced" *Optimal Transport* and its variant for *Unbalanced Optimal Transport*. We further show how it is connected to the softmax.

C.1. Sinkhorn and Variant

These two algorithms are taken from [6, 2]. In particular we can see how taking $\tau_1 \to +\infty$ and $\tau_2 \to +\infty$ in Algorithm 2 leads to Algorithm 1. Indeed, we have $\lim_{\tau \to +\infty} \frac{\tau}{\tau+\epsilon} = 1$. By $\oslash$, we denote the element-wise (or Hadamard) division.

```
Data: Distributions $\boldsymbol{\alpha} \in \Delta^{N_p}$ and $\boldsymbol{\beta} \in \Delta^{N_g+1}$, regularization parameter $\epsilon \in \mathbb{R}_{>0}$ and cost matrix
      $\boldsymbol{C} = \left[\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right]_{i,j=1}^{N_p,N_g+1} \in \mathbb{R}_{\geq 0}^{N_p \times N_g+1}$ (including background $\boldsymbol{y}_{N_g+1} = \varnothing$).
Result: Match $\hat{\boldsymbol{P}} \in \Delta^{N_p,N_g+1}$.
begin
    $\boldsymbol{K}_\epsilon \longleftarrow \exp\left(-\boldsymbol{C}/\epsilon\right)$            /* Gramm matrix (element-wise) */
    $\boldsymbol{u} \longleftarrow \mathbf{1}_{N_p}/N_p$                  /* Dual variable associated with $\boldsymbol{\alpha}$ */
    $\boldsymbol{v} \longleftarrow \mathbf{1}_{N_g+1}/\left(N_g+1\right)$         /* Dual variable associated with $\boldsymbol{\beta}$ */
    repeat
        $\boldsymbol{u} \longleftarrow \boldsymbol{\alpha} \oslash \left(\boldsymbol{K}_\epsilon \boldsymbol{v}\right)$          /* Scaling iteration for $\boldsymbol{u}$ */
        $\boldsymbol{v} \longleftarrow \boldsymbol{\beta} \oslash \left(\boldsymbol{K}_\epsilon^\top \boldsymbol{u}\right)$         /* Scaling iteration for $\boldsymbol{v}$ */
    until convergence
    $\hat{\boldsymbol{P}} \longleftarrow \boldsymbol{u}\boldsymbol{K}_\epsilon\boldsymbol{v}$
```

Algorithm 1: Sinkhorn's algorithm for "balanced" *Optimal Transport* with regularization.

```
Data: Distributions $\boldsymbol{\alpha} \in \Delta^{N_p}$ and $\boldsymbol{\beta} \in \Delta^{N_g+1}$, regularization parameter $\epsilon \in \mathbb{R}_{>0}$, constraint parameters
      $\tau_1, \tau_2 \in \mathbb{R}_{\geq 0}$ and cost matrix $\boldsymbol{C} = \left[\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right]_{i,j=1}^{N_p,N_g+1} \in \mathbb{R}_{\geq 0}^{N_p \times N_g+1}$ (including background
      $\boldsymbol{y}_{N_g+1} = \varnothing$).
Result: Match $\hat{\boldsymbol{P}} \in \mathbb{R}_{\geq 0}^{N_p,N_g+1}$.
begin
    $\boldsymbol{K}_\epsilon \longleftarrow \exp\left(-\boldsymbol{C}/\epsilon\right)$            /* Gramm matrix (element-wise) */
    $\boldsymbol{u} \longleftarrow \mathbf{1}_{N_p}/N_p$                  /* Dual variable associated with $\boldsymbol{\alpha}$ */
    $\boldsymbol{v} \longleftarrow \mathbf{1}_{N_g+1}/\left(N_g+1\right)$         /* Dual variable associated with $\boldsymbol{\beta}$ */
    repeat
        $\boldsymbol{u} \longleftarrow \left(\boldsymbol{\alpha} \oslash \left(\boldsymbol{K}_\epsilon \boldsymbol{v}\right)\right)^{\frac{\tau_1}{\tau_1+\epsilon}}$          /* Scaling iteration for $\boldsymbol{u}$ */
        $\boldsymbol{v} \longleftarrow \left(\boldsymbol{\beta} \oslash \left(\boldsymbol{K}_\epsilon^\top \boldsymbol{u}\right)\right)^{\frac{\tau_2}{\tau_2+\epsilon}}$         /* Scaling iteration for $\boldsymbol{v}$ */
    until convergence
    $\hat{\boldsymbol{P}} \longleftarrow \boldsymbol{u}\boldsymbol{K}_\epsilon\boldsymbol{v}$
```

Algorithm 2: Scaling algorithm for *Unbalanced Optimal Transport* with regularization.

C.2. Connection with the Softmax

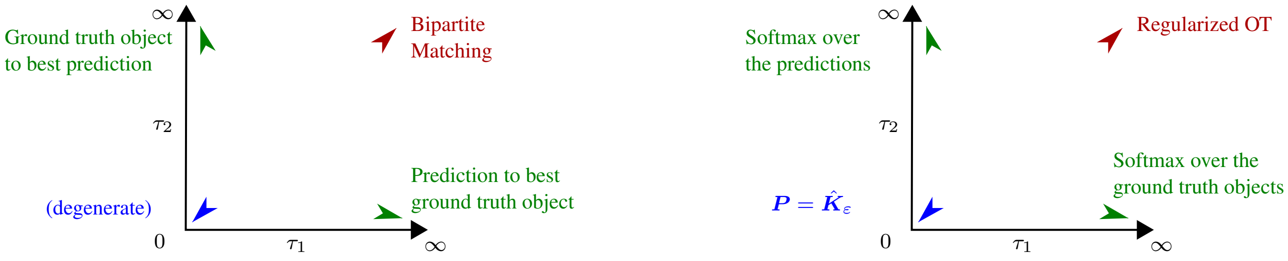
In this section, we lay a connection between the softmax and the solutions of the scaling algorithms, in particular considering its first iterations. We consider more precisely the softmin, which is the opposite of the softmax: $\left(\text{softmin}(\boldsymbol{v})\right)_i = \left(\text{softmax}(-\boldsymbol{v})\right)_i = \exp(-v_i)/\sum_{j=1}^{N}\exp(-v_j)$, for any vector $\boldsymbol{v} \in \mathbb{R}^N$. Considering a softmin over $\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)$ is thus the same as considering the softmax over $-\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)$, as in [8]. By simplicity, we will only use the softmax terminology.

C.2.1 Without Background

We first consider the case without background, where the underlying distributions are equal to $\boldsymbol{\alpha} = \mathbf{1}_{N_p}/N_p$ and $\boldsymbol{\beta} = \mathbf{1}_{N_g}/N_g$. This does not correspond to the setup of Prop. 1 and only approximates a one-to-one match if $N_p = N_g$.

Proposition 3. *Consider the two uniform distributions $\boldsymbol{\alpha} = \mathbf{1}_{N_p}/N_p$ and $\boldsymbol{\beta} = \mathbf{1}_{N_g}/N_g$ with cost $\mathcal{L}_{match}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)$. The solution of the Unbalanced OT scaling algorithm with regularization $\varepsilon = 1$, $\tau_1 = 0$ and $\tau_2 \to +\infty$ is proportional to performing a softmax over the predictions, for each ground truth object. In particular, we have*

$$\hat{P}_{i,j} = \frac{\exp\left(-\mathcal{L}_{match}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right)}{N_g \sum_{i=1}^{N_p} \exp\left(-\mathcal{L}_{match}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right)}. \tag{18}$$


∞
Ground truth object to best prediction
Bipartite Matching
τ_2
Prediction to best ground truth object
(degenerate)
0
τ_1
∞
∞
Softmax over the predictions
Regularized OT
τ_2
Softmax over the ground truth objects
$\boldsymbol{P} = \hat{\boldsymbol{K}}_\varepsilon$
0
τ_1
∞

(a) No regularization $\varepsilon = 0$ (same as Fig. 4).

(b) With regularization $\varepsilon = 1$.

Figure 7. Comparison of the different limit cases of *Unbalanced Optimal Transport*, with and without regularization.

Proof. We consider the first scaling iteration from Alg. 2. We first observe that the exponents lead to $\lim_{\tau_1 \to 0} \frac{\tau_1}{\tau_1 + \varepsilon} = 0$ and $\lim_{\tau_2 \to +\infty} \frac{\tau_2}{\tau_2 + \varepsilon} = 1$. Starting with $\boldsymbol{v}_0 = \mathbf{1}_{N_g}/N_g$, we obtain the new

$$\boldsymbol{u}_1 = \left(\boldsymbol{\alpha} \oslash (\boldsymbol{K}_\varepsilon \boldsymbol{v}_0)\right)^0 = \mathbf{1}_N, \qquad \text{or } (\boldsymbol{u}_1)_i = 1, \tag{19}$$

$$\boldsymbol{v}_1 = \left(\boldsymbol{\beta} \oslash (\boldsymbol{K}_\varepsilon^\top \boldsymbol{u}_1)\right)^1 = \left(\mathbf{1}_{N_g}/N_g\right) \oslash \left(\boldsymbol{K}_\varepsilon^\top \mathbf{1}_N\right), \qquad \text{or } (\boldsymbol{v}_1)_j = \frac{1}{N_g \sum_{i=1}^{N_p} \exp\left(-\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right)}. \tag{20}$$

We observe that $\boldsymbol{u}_2 = \boldsymbol{u}_1$ and $\boldsymbol{v}_2 = \boldsymbol{v}_1$ and conclude that the algorithm converges after only one iteration. Computing the match $\hat{\boldsymbol{P}} = \boldsymbol{u}\boldsymbol{K}_\epsilon\boldsymbol{v}$ leads to the softmax. □

The exact opposite happens if we consider $\tau_1 \to +\infty$ and $\tau_2 = 0$ instead: the softmax is taken over the ground truth objects for each prediction. The proof is the same, just inverting $\boldsymbol{u}$ and $\boldsymbol{v}$ and obtaining factor $1/N_p$ instead. This can be observed at Fig. 7b.

If we would like to exactly obtain the softmax without the factor $1/N_g$ (or $1/N_p$), we could consider only one iteration starting with both initial dual variables $\boldsymbol{u}_0$ and $\boldsymbol{v}_0$. It would however not be the optimal match $\hat{\boldsymbol{P}}$ and will converge to the same solution as in Prop. 3 after the second—and last—iteration. Nevertheless, starting from both initial dual variables is more interesting in the "balanced" case.

Proposition 4. *Consider the two uniform distributions $\boldsymbol{\alpha} = \mathbf{1}_{N_p}/N_p$ and $\boldsymbol{\beta} = \mathbf{1}_{N_g}/N_g$ with cost $\mathcal{L}_{match}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)$. Starting from both initial dual variables, one iteration of the "balanced" OT scaling algorithm with regularization $\varepsilon = 1$ is equal to*

$$P_{i,j} = \frac{\exp\left(-\mathcal{L}_{match}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right)}{\sum_{i=1}^{N_p} \exp\left(-\mathcal{L}_{match}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right) \cdot \sum_{j=1}^{N_g} \exp\left(-\mathcal{L}_{match}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right)}. \tag{21}$$

Proof. We consider the first scaling iteration from Alg. 1 with $\boldsymbol{\alpha} = \mathbf{1}_{N_p}/N_p$ and $\boldsymbol{\beta} = \mathbf{1}_{N_g}/N_g$. Starting with $\boldsymbol{u}_0 = \mathbf{1}_{N_p}/N_p$ and $\boldsymbol{v}_0 = \mathbf{1}_{N_g}/N_g$, we obtain the new

$$\boldsymbol{u}_1 = \boldsymbol{\alpha} \oslash (\boldsymbol{K}_\varepsilon \boldsymbol{v}_0) = \left(\mathbf{1}_{N_p}/N_p\right) \oslash \left(\boldsymbol{K}_\varepsilon \left(\mathbf{1}_{N_g}/N_g\right)\right), \qquad \text{or } (\boldsymbol{u}_1)_i = \frac{N_g}{N_p \sum_{j=1}^{N_g} \exp\left(-\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right)}, \tag{22}$$

$$\boldsymbol{v}_1 = \boldsymbol{\beta} \oslash (\boldsymbol{K}_\varepsilon^\top \boldsymbol{u}_0) = \left(\mathbf{1}_{N_g}/N_g\right) \oslash \left(\boldsymbol{K}_\varepsilon^\top \left(\mathbf{1}_{N_p}/N_p\right)\right), \qquad \text{or } (\boldsymbol{v}_1)_j = \frac{N_p}{N_g \sum_{i=1}^{N_p} \exp\left(-\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right)}. \tag{23}$$

Computing the match $\boldsymbol{P} = \boldsymbol{u}\boldsymbol{K}_\epsilon\boldsymbol{v}$ leads to the Eq. 21. This is not the optimal match $\hat{\boldsymbol{P}}$ as the algorithm did not converge yet. □

The *dual-softmax* considered in [8] is essentially the same as Prop. 4, with the difference of a factor 2 in the numerator's exponential:

$$\begin{aligned} P_{i,j} &= \operatorname{softmax}\left(\left[-\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_k\right)\right]_{k=1}^{N_g}\right)_j \cdot \operatorname{softmax}\left(\left[-\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_l, \boldsymbol{y}_j\right)\right]_{j=1}^{N_p}\right)_i, &(24)\\ &= \frac{\exp\left(-2\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right)}{\sum_{i=1}^{N_p} \exp\left(-\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right) \cdot \sum_{j=1}^{N_g} \exp\left(-\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right)}. &(25) \end{aligned}$$

C.2.2 With Background

We now consider the underlying distributions as defined in Prop. 1. Fundamentally, adding a background with a different weight than the other ground truth objects does not change much. The unbalanced case with $\tau_1 \to +\infty$ and $\tau_2 = 0$ remains exactly the same. The opposite case with $\tau_1 = 0$ and $\tau_2 \to +\infty$ now becomes

$$\hat{P}_{i,j} = \frac{1}{N_p} \frac{\exp\left(-\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right)}{\sum_{i=1}^{N_p} \exp\left(-\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right)}, \tag{26}$$

for all $1 \leq j \leq N_g$, and

$$\hat{P}_{i,j} = \frac{N_p - N_g}{N_p} \frac{\exp\left(-\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right)}{\sum_{i=1}^{N_p} \exp\left(-\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right)}, \tag{27}$$

for $j = N_g + 1$ (the background). In essence, this ensures that the mass constraints induced by τ_2 are satisfied, as the background has a higher weight.

Similarly, the "balanced" case is the same as Eq. 21 for all $1 \leq j \leq N_g$. For $j = N_g + 1$, we have the same with an added factor:

$$P_{i,j} = (N_p - N_g) \frac{\exp\left(-\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right)}{\sum_{i=1}^{N_p} \exp\left(-\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right) \cdot \sum_{j=1}^{N_g} \exp\left(-\mathcal{L}_{\text{match}}\left(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j\right)\right)}. \tag{28}$$

C.2.3 Other Regularization

We can also consider other cases that having the regularization $\varepsilon = 1$. The regularization ε controls the "softness" of the softmax: the greater is ε, the softer is the minimum; the smaller, the harder. In the case of no regularization at all ($\varepsilon \to 0$), the softmax is exactly a minimum as proven in Prop. 2. This can be observed at Fig. 8.

D. Scaling the Entropic Parameter

In this section, we consider the particular choice of the entropic regularization parameter. In particular, we study how it scales with the problem size.

D.1. Uniform Matches

Definition 6 (Matches). *We define a* match $\boldsymbol{P} \in \mathbb{R}_+^{N_p \times (N_g+1)}$ *as a positive matrix of unity mass* $\sum_{i,j} P_{i,j} = 1$. *The set of all matches of size* $N_p \times (N_g + 1)$ *is the joint probability simplex* $\Delta^{N_p \times (N_g+1)}$.

We now consider a particular subset of all these matches.

Definition 7 (Uniform Matches). *We define the set of* uniform matches $\Delta_{\text{unif.}}^{N_p \times (N_g+1)} \subsetneq \Delta^{N_p \times (N_g+1)}$ *as the set of matrices* $\boldsymbol{P}^{\text{unif.}} \in \Delta_{\text{unif.}}^{N_p \times (N_g+1)}$, *containing only zero elements and all non-zero elements having the same value:*

$$P_{i,j}^{\text{unif.}} = \begin{cases} 0 & \textit{for some values,} \\ 1/\left|\operatorname{spt}\left(\boldsymbol{P}^{\text{unif.}}\right)\right| & \textit{for the other values,} \end{cases} \tag{29}$$

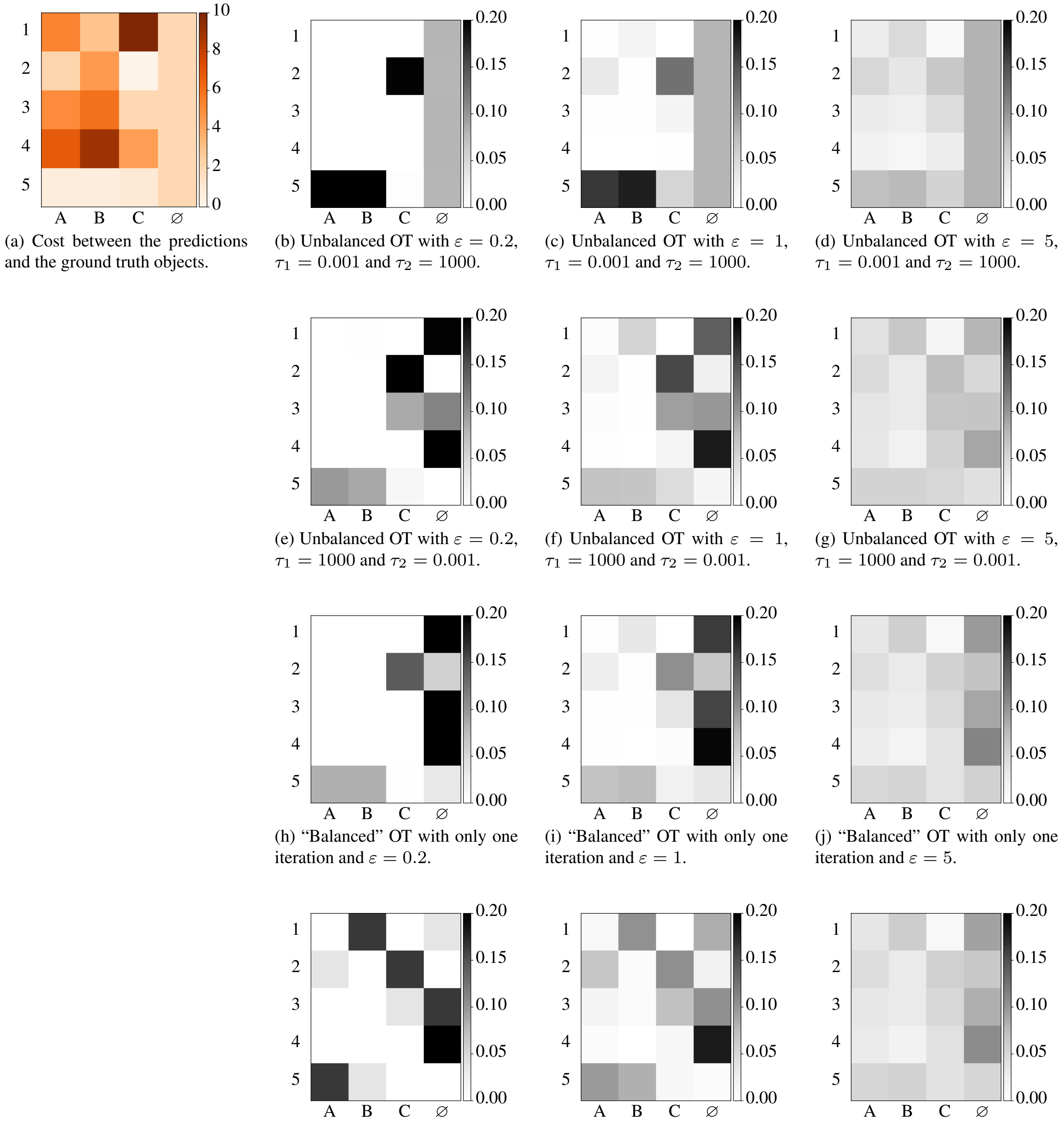
with the support $\operatorname{spt} : \boldsymbol{P} \mapsto \{(i,j) : P_{i,j} \neq 0\}$ *and* $|\cdot|$ *the cardinality of a set.*

We directly see from the definition that the matrices are well defined as they have unity mass. They are uniquely defined by the carnality of their support.

Proposition 5 (Cardinality). *The cardinality of* $\Delta_{\text{unif.}}^{N_p \times (N_g+1)}$ *is given by*

$$\left|\Delta_{\text{unif.}}^{N_p \times (N_g+1)}\right| = 2^{N_p (N_g+1)}. \tag{30}$$

Proof. We first notice that the different possible supports $k = \operatorname{spt}\left(\boldsymbol{P}^{\text{unif.}}\right)$ range from $1 \leq k \leq N_p\left(N_g + 1\right)$. For any support of size k, we have to consider a uniform match containing all combinations. The rest follows from the binomial identity $\sum_{k=1}^{N_p(N_g+1)} \binom{N_p(N_g+1)}{k} = 2^{N_p(N_g+1)}$. □



(a) Cost between the predictions and the ground truth objects.

(b) Unbalanced OT with $\varepsilon = 0.2$, $\tau_1 = 0.001$ and $\tau_2 = 1000$.

(c) Unbalanced OT with $\varepsilon = 1$, $\tau_1 = 0.001$ and $\tau_2 = 1000$.

(d) Unbalanced OT with $\varepsilon = 5$, $\tau_1 = 0.001$ and $\tau_2 = 1000$.

(e) Unbalanced OT with $\varepsilon = 0.2$, $\tau_1 = 1000$ and $\tau_2 = 0.001$.

(f) Unbalanced OT with $\varepsilon = 1$, $\tau_1 = 1000$ and $\tau_2 = 0.001$.

(g) Unbalanced OT with $\varepsilon = 5$, $\tau_1 = 1000$ and $\tau_2 = 0.001$.

(h) “Balanced” OT with only one iteration and $\varepsilon = 0.2$.

(i) “Balanced” OT with only one iteration and $\varepsilon = 1$.

(j) “Balanced” OT with only one iteration and $\varepsilon = 5$.

(k) “Balanced” OT until convergence with $\varepsilon = 0.2$.

(l) “Balanced” OT until convergence with $\varepsilon = 1$.

(m) “Balanced” OT until convergence with $\varepsilon = 5$.

Figure 8. Connection between scaling algorithms and the softmax. The pairwise matching cost between the predictions (numbers) and the ground truth objects (letters) is given in Fig. 8a. The background cost is $c_\varnothing = 2$. The scaling algorithm for Unbalanced OT corresponds to performing the softmax column-wise (Figs. 8b, 8c and 8d), or row-wise (Figs. 8e, 8f and 8g). Similarly, one iteration of the scaling algorithm for “balanced” OT is almost equivalent to the dual-softmax (Figs. 8h, 8i and 8j), but does not satisfy the mass constraints unlike when it is run until convergence (Figs. 8k, 8l and 8m).

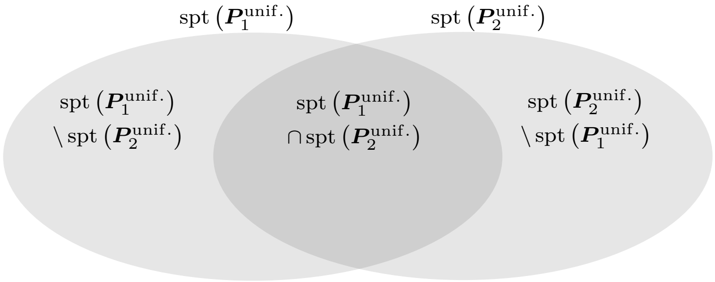

$\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right)$
$\operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right)$
$\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right)$
$\setminus \operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right)$
$\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right)$
$\cap \operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right)$
$\operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right)$
$\setminus \operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right)$

Figure 9. Decomposition of the non-zero indices of two uniform transport matrices $\boldsymbol{P}_1^{\text{unif.}}, \boldsymbol{P}_2^{\text{unif.}} \in \Delta_{\text{unif.}}^{N_p \times \left(N_g+1\right)}$.

We can also see that the uniform matches cover the set of all matches.

Proposition 6 (Diameter)**.** *The diameter of the set of transport matrices $\Delta^{N_p \times (N_g+1)}$ and uniform transport matrices $\Delta_{\text{unif.}}^{N_p \times (N_g+1)}$, equipped with the Fröbenius norm $\| \cdot \|_F$, is given by*

$$\operatorname{diam}\left(\Delta^{N_p \times (N_g+1)}\right) = \operatorname{diam}\left(\Delta_{\text{unif.}}^{N_p \times (N_g+1)}\right) = \sqrt{2} \tag{31}$$

Proof. Maximizing the Fröbenius norm is equivalent to considering the maximization of $\sum_i (u_i - v_i)^2$ subject to $\sum_i u_i = 1$ and $\sum_i v_i = 1$, with $\boldsymbol{u}, \boldsymbol{v} \geq 0$. It takes its maximum value on the boundary of the admissible solutions, for $u_i = 1$ (the rest is zero) and $v_j = 1$ (the rest zero) for any $j \neq i$. These extreme points are also in $\Delta_{\text{unif.}}^{N_p \times (N_g+1)}$, in particular those of unity support $\operatorname{spt}\left(\boldsymbol{P}^{\text{unif.}}\right) = 1$. □

Proposition 7. *The Fröbenius norm square $\|\boldsymbol{P}_1^{\text{unif.}} - \boldsymbol{P}_2^{\text{unif.}}\|_F^2$ between two uniform matches $\boldsymbol{P}_1^{\text{unif.}}, \boldsymbol{P}_2^{\text{unif.}} \in \Delta_{\text{unif.}}^{N_p \times (N_g+1)}$ is given by*

$$\frac{\left|\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right)\right| + \left|\operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right)\right| - 2\left|\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right) \cap \operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right)\right|}{\left|\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right)\right| \left|\operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right)\right|}. \tag{32}$$

Proof. By decomposing the all indices as in Figure 9 in

$$\begin{aligned}
\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right) \cup \operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right) \quad = \quad & \left(\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right) \setminus \operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right)\right) \\
& \cup \left(\operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right) \setminus \operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right)\right) \\
& \cup \left(\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right) \cap \operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right)\right),
\end{aligned}$$

and noticing that all other values are zero, we have for $\|\boldsymbol{P}_1^{\text{unif.}} - \boldsymbol{P}_2^{\text{unif.}}\|_F^2$

$$\begin{aligned}
& \left(\left|\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right)\right| - \left|\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right) \cap \operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right)\right|\right) \frac{1}{\left|\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right)\right|^2} \\
+ \quad & \left(\left|\operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right)\right| - \left|\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right) \cap \operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right)\right|\right) \frac{1}{\left|\operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right)\right|^2} \\
+ \quad & \left|\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right) \cap \operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right)\right| \left(\frac{1}{\left|\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right)\right|} - \frac{1}{\left|\operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right)\right|}\right)^2.
\end{aligned}$$

The rest is just a simplification of the latter. □

Corollary 3. *Each uniform match $\boldsymbol{P}_1^{\text{unif.}} \in \Delta_{\text{unif.}}^{N_p \times (N_g+1)}$ has as closest neighbors all other uniform matches $\boldsymbol{P}_2^{\text{unif.}} \in \Delta_{\text{unif.}}^{N_p \times (N_g+1)}$ of support increased by one $\left|\operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right)\right| = \left|\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right)\right| + 1$ and differing in support for only one entry $\left|\operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right) \setminus \operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right)\right| = 1$. In particular, the square Fröbenius norm is then equal to*

$$\left\|\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right) - \operatorname{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right)\right\|_F^2 = \frac{1}{\left|\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right)\right| \left(\left|\operatorname{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right)\right| + 1\right)}. \tag{33}$$

In in the particular limit case of $\left|\mathrm{spt}\left(\boldsymbol{P}_1^{\text{unif.}}\right)\right| = N_p\left(N_g+1\right)$*, its closest neighbors are all the* $\boldsymbol{P}_2^{\text{unif.}}$ *such that* $\left|\mathrm{spt}\left(\boldsymbol{P}_2^{\text{unif.}}\right)\right| = N_p\left(N_g+1\right)-1$.

Proposition 8. *We consider the projector* $\mathbb{P} : \Delta^{N_p\times(N_g+1)} \to \Delta_{\text{unif.}}^{N_p\times(N_g+1)}$*, that minimizes the Fröbenius norm. For any* $\boldsymbol{P}\in\Delta^{N_p\times(N_g+1)}$*, we consider*

$$\mathbb{P}(\boldsymbol{P}) = \underset{\boldsymbol{P}^{\text{unif.}}\in\Delta_{\text{unif.}}^{N_p\times(N_g+1)}}{\mathrm{argmin}} \|\boldsymbol{P}-\boldsymbol{P}^{\text{unif.}}\|_F. \tag{34}$$

It is given by the matrix $\boldsymbol{P}^{\text{unif.}}\in\Delta_{\text{unif.}}^{N_p\times(N_g+1)}$ *with the* k *greatest elements of* $\boldsymbol{P}$ *as support and*

$$k = \underset{k\in[\![N_p(N_g+1)]\!]}{\arg\max}\ \frac{1}{k}\left(2\sum_{\substack{k\text{ greatest}\\\text{elements}}} P_{ij}-1\right). \tag{35}$$

Proof. We consider the distance between any element $\boldsymbol{P}\in\Delta^{N_p\times(N_g+1)}$ and $\boldsymbol{P}^{\text{unif.}}\in\Delta_{\text{unif.}}^{N_p\times(N_g+1)}$: $\|\boldsymbol{P}-\boldsymbol{P}^{\text{unif.}}\|_F^2 = \sum_{i,j=1}^{N_p,(N_g+1)}\left(P_{i,j}-P_{i,j}^{\text{unif.}}\right)^2 = \sum_{i,j=1}^{N_p,(N_g+1)} P_{i,j}^2+\left(P_{i,j}^{\text{unif.}}\right)^2-2P_{ij}P_{i,j}^{\text{unif.}}$. We notice that because of the uniform nature of $\boldsymbol{P}^{\text{unif.}}$, it only has k non-zero elements, all equal to $1/k$. By consequence we have $\sum_{i,j=1}^{N_p,(N_g+1)}\left(P_{ij}^{\text{unif.}}\right)^2 = \frac{1}{k}$ and $\sum_{i,j=1}^{N_p,(N_g+1)} P_{ij}P_{ij}^{\text{unif.}} = \frac{1}{k}\sum_{\mathrm{spt}(\boldsymbol{P}^{\text{unif.}})} P_{ij}$. The distance is now equal to $\|\boldsymbol{P}-\boldsymbol{P}^{\text{unif.}}\|_F^2 = \|\boldsymbol{P}\|_F^2+\frac{1}{k}-\frac{2}{k}\sum_{\mathrm{spt}(\boldsymbol{P}')} P_{ij}$, and is minimal if $\frac{2}{k}\sum_{\mathrm{spt}(\boldsymbol{P}^{\text{unif.}})} P_{ij}-\frac{1}{k}$ is maximal which is unique as it suffices to see that is it reached once

$$\sum_{\substack{k\text{ greatest}\\\text{elements}}} P_{ij} > P_{\substack{\text{next}\\\text{greatest}}}+\frac{1}{2}, \tag{36}$$

is satisfied. □

The norm with the projector is therefore also given by $\|\boldsymbol{P}-\mathbb{P}(\boldsymbol{P})\|_F^2 = \|\boldsymbol{P}\|_F - 2\sum_{\mathrm{spt}(\mathbb{P}(\boldsymbol{P}))} P_{ij}+\frac{1}{|\mathrm{spt}(\mathbb{P}(\boldsymbol{P}))|}$. Equation (36) gives a direct algorithm to determine k and thus the projected value of any match $\boldsymbol{P}$.

D.2. Entropy

The study of uniform matches is relevant as they have an easy formulation for their entropy.

Definition 8 (Entropy)**.** *The entropy* $\mathrm{H}:\Delta^{N_p\times(N_g+1)}\to\mathbb{R}_{\geq 0}$ *of a match* $\boldsymbol{P}$ *is given by*

$$\mathrm{H}(\boldsymbol{P}) \overset{\text{def.}}{=} -\sum_{i,j} P_{i,j}\left(\log\left(P_{i,j}\right)-1\right). \tag{37}$$

If one of the elements would be zero, i.e., $P_{i,j}=0$*, we consider* $P_{i,j}\log\left(P_{i,j}-1\right)=0$.

The latter condition ensures that the entropy is well defined. This choice is justified as it remains consistent with the limit. Some authors prefer another convention [6].

Lemma 3. *The entropy of a uniform match* $\boldsymbol{P}^{\text{unif.}}\in\Delta_{\text{unif.}}^{N_p\times(N_g+1)}$ *is given by*

$$\mathrm{H}(\boldsymbol{P}^{\text{unif.}}) = \log\left(\left|\mathrm{spt}(\boldsymbol{P}^{\text{unif.}})\right|\right)+1. \tag{38}$$

Proof. The proof is a direct application of the definition of the entropy:

$$\begin{aligned}
\mathrm{H}(\boldsymbol{P}^{\text{unif.}}) &= -\sum_{i,j} P_{i,j}^{\text{unif.}}\left(\log\left(P_{i,j}^{\text{unif.}}\right)-1\right), && (39)\\
&= -\sum_{\mathrm{spt}(\boldsymbol{P}^{\text{unif.}})} P_{i,j}^{\text{unif.}}\left(\log\left(P_{i,j}^{\text{unif.}}\right)-1\right), && (40)\\
&= -\frac{\left|\mathrm{spt}\left(\boldsymbol{P}^{\text{unif.}}\right)\right|}{\left|\mathrm{spt}\left(\boldsymbol{P}^{\text{unif.}}\right)\right|}\left(\log\left(\frac{1}{\left|\mathrm{spt}\left(\boldsymbol{P}^{\text{unif.}}\right)\right|}\right)-1\right), && (41)\\
&= \log\left(\left|\mathrm{spt}(\boldsymbol{P}^{\text{unif.}})\right|\right)+1. && (42)
\end{aligned}$$

□

Proposition 9. *For any match $\boldsymbol{P} \in \Delta^{N_p \times (N_g+1)}$,*

$$1 \leq H(\boldsymbol{P}) \leq \log(N_p (N_g + 1)) + 1. \tag{43}$$

Proof. For an arbitrary coupling matrix, the entropy is always minimal if $P_{i,j} = 1$ for one element and all the others are zero. Similarly, the entropy is always for the uniform match $P_{i,j} = 1/\left|\operatorname{spt}(\boldsymbol{P})\right|$ for all i, j, with $\left|\operatorname{spt}(\boldsymbol{P})\right| = N_p \times (N_g + 1)$. □

D.3. Rule of Thumb

We first consider two different matches of different dimensions $\boldsymbol{P}_1 \in \Delta^{N_{p,1} \times (N_{g,1}+1)}$ and $\boldsymbol{P}_2 \in \Delta^{N_{p,2} \times (N_{g,2}+1)}$. In this case, the OT with regularization cost (Definition 3) is given by $\sum_{i,j=1}^{N_p,(N_g+1)} P_{i,j} \mathcal{L}_{\text{match}}(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j) - \epsilon\,\mathrm{H}(\boldsymbol{P})$. The goal is to scale the regularization parameter ϵ in such a way that the weight of the entropy is proportionally the same. Because of unit mass of any match, we could assume that the first term $\sum_{i,j=1}^{N_p,(N_g+1)} P_{i,j} \mathcal{L}_{\text{match}}(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j)$ is independent of N_p and N_g in magnitude. We therefore have to guarantee that $\epsilon_1 \mathrm{H}(\boldsymbol{P}_1) = \epsilon_2 \mathrm{H}(\boldsymbol{P}_2)$. Given an already determined regularization value ϵ_1 for one of the two sizes, the other can be found with $\epsilon_2 = \epsilon_1 \mathrm{H}(\boldsymbol{P}_1) / \mathrm{H}(\boldsymbol{P}_2)$. In practice, however, the entropy is not trivial and we can rely on the projection onto the uniform matches

$$\epsilon_1 = \epsilon_2 \frac{\log\left(\left|\operatorname{spt}\left(\mathbb{P}\left(\boldsymbol{P}_2\right)\right)\right|\right) + 1}{\log\left(\left|\operatorname{spt}\left(\mathbb{P}\left(\boldsymbol{P}_1\right)\right)\right|\right) + 1}. \tag{44}$$

In the particular case of Proposition 1, we can use the approximation $\left|\operatorname{spt}\left(\mathbb{P}\left(\boldsymbol{P}\right)\right)\right| = N_p$, which gives

$$\epsilon_1 = \epsilon_2 \frac{\log\left(N_{p,2}\right) + 1}{\log\left(N_{p,1}\right) + 1}. \tag{45}$$

The idea is to determine the optimal value ϵ_1 on toy examples. By setting $N_p = N_{p,2}$, $\epsilon = \epsilon_2$ and $\epsilon_0 = \epsilon_1 \left(\log\left(N_{p,1}\right) + 1\right)$, we can use the simple scaling formula $\epsilon = \epsilon_0 / \left(\log\left(N_p\right) + 1\right)$. From our experiments, we determined $\epsilon_0 = 0.12$.

E. Qualitative Analysis

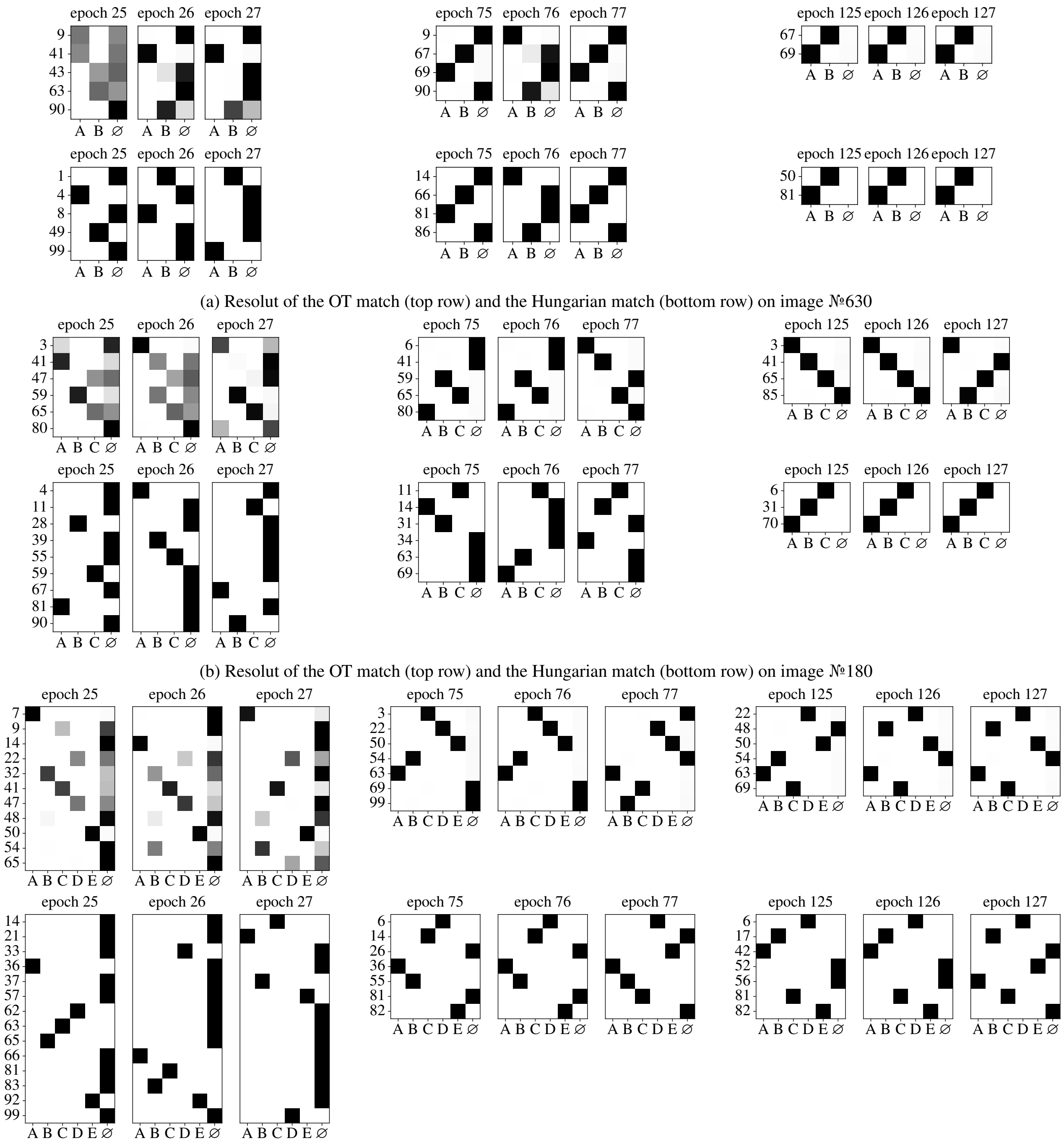
This section provides qualitative examples (Figure 10 and Figure 11) of some matches, as well as a convergence analysis for DETR and Deformable DETR. We compare the losses and matches $\boldsymbol{P}$ of the two matching algorithms at different training epochs.

Figure 10 shows some assignments of the two matching algorithms for DETR on the Color Boxes dataset. We sample examples with few ground truth objects for readability. We only show predictions that are matched at least once with a background $\varnothing$ ground truth in three consecutive epochs. At the beginning of the training, the *Bipartite Matching* with the *Hungarian algorithm* assigns different predictions to the ground truth objects from one epoch to the other. As an example, the algorithm for image №630 assigns predictions $\{4, 49\}$, $\{8, 1\}$ and then $\{99, 1\}$ to the ground-truth objects $\{A, B\}$ at epoch 25 to 27 (Figure 10a). The regularized OT match instead provides a smoother solution and is more consistent from one epoch to the other. Later in training, Figure 10a illustrates that the regularized OT matches are one-to-one and behave like the bipartite ones.

Figure 12 provides the loss curves for DETR on the Color Boxes dataset. The curves suggest that the cross-entropy loss term mainly drives the convergence speedup in the early training epochs. We don't observe such speedups on COCO or with Deformable DETR (Figure 13). An explanation could be that the difference between DETR and Deformable DETR is due to the slower convergence of transformers (we also tried DETR with the focal loss from Deformable DETR without improvement). The difference between Color Boxes and COCO is difficult to isolate, but probably due to the wider class diversity in the latter.

F. Number of Sinkhorn Iterations

Using a stopping criterion is not straightforward when solving a batch of matching problems. The scaling algorithm is therefore set to a fixed number of iterations. Figure 14 displays the results for different numbers of iterations. For the balanced OT with 300 predictions (Figure 14a), the AP increases only slightly when more than 10 iterations are performed. Furthermore, it is sufficient to run 1 iteration in terms of the AR. For the *Unbalanced OT* with 8,732 predictions (Figure 14b), the metrics are significantly lower when running for less than 5 iterations. Again, running more than 10 iterations only slightly improves the final performance. This fact is supported by Prop. 3, which shows that in the limit case where $\tau_1 = 0$ and $\tau_2 \to +\infty$, only one or two iterations are required for convergence (depending on the implementation).



(a) Resolut of the OT match (top row) and the Hungarian match (bottom row) on image №630

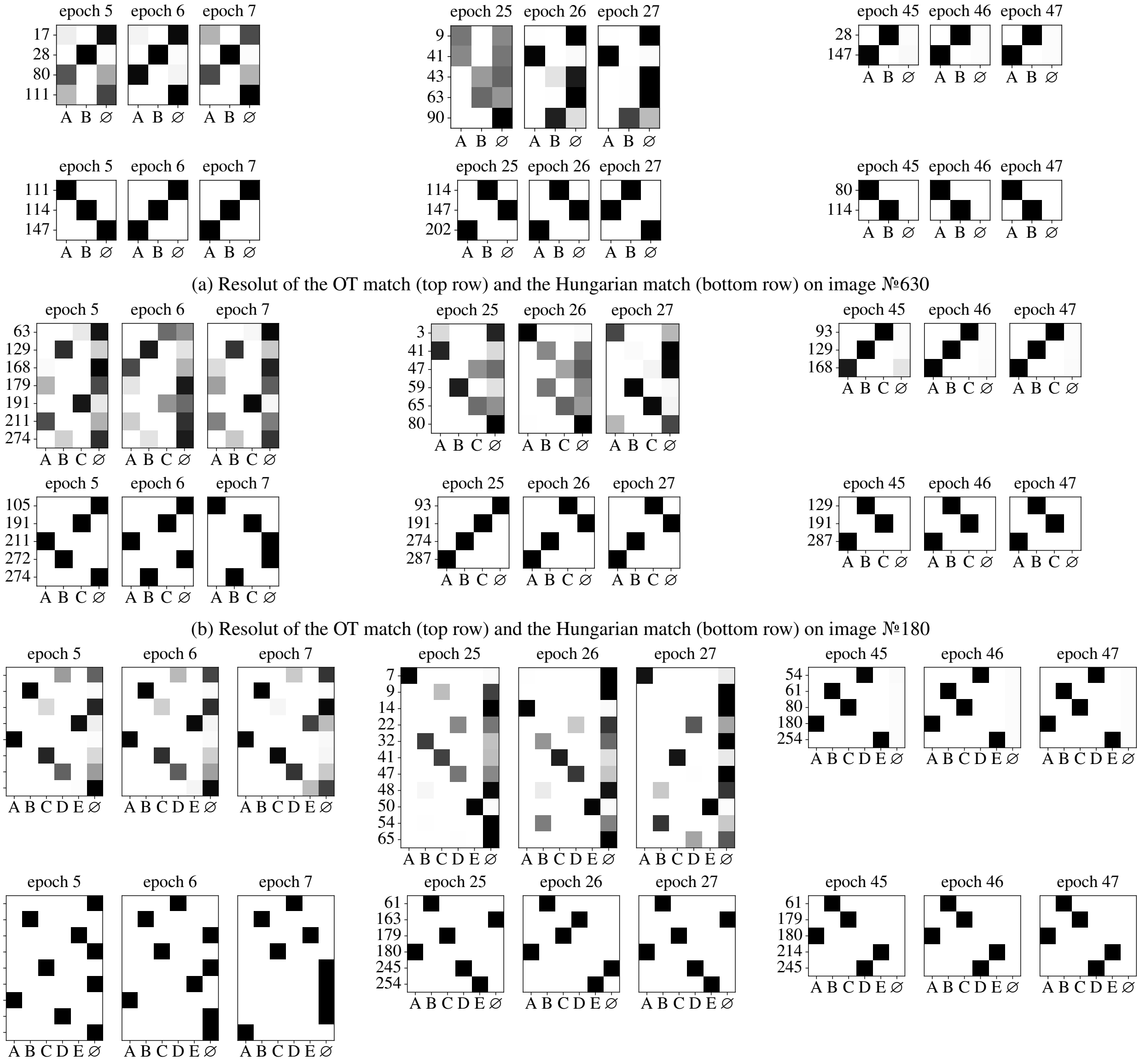
(b) Resolut of the OT match (top row) and the Hungarian match (bottom row) on image №180

(c) Resolut of the OT match (top row) and the Hungarian match (bottom row) on image №613

Figure 10. Output of the matching algorithms with DETR on the validation set of the Color Boxes Dataset. The model is trained two times: once with an OT match and once with a Hungarian matching. The rows indicate the predictions and the columns indicate the ground truth objects (including the background ∅). We sample examples with few ground truth objects for readability and only show predictions that are matched at least once with a non-background ground truth.

G. First Constraint Parameter

In this section, we analyze the effect of the prediction's mass constraint parameter τ_1, while we fix parameter τ_2 to a large value $\tau_2 = 100$ to simulate a hard constraint. Parameter τ_1 controls the degree to which variations in the prediction masses



(a) Resolut of the OT match (top row) and the Hungarian match (bottom row) on image №630

(b) Resolut of the OT match (top row) and the Hungarian match (bottom row) on image №180

(c) Resolut of the OT match (top row) and the Hungarian match (bottom row) on image №613

Figure 11. Output of the matching algorithms with Deformable-DETR on the validation set of the Color Boxes Dataset. The model is trained two times: once with an OT match and once with a Hungarian matching. The rows indicate the predictions and the columns indicate the ground truth objects (including the background ∅). We sample examples with few ground truth objects for readability and only show predictions that are matched at least once with a non-background ground truth.

are penalized. Each ground-truth object can be matched to the best prediction in the limit case $\tau_2 \to +\infty$ and $\tau_1 = 0$. However, some predictions cannot be matched, and others multiple times. The results for SSD on Color Boxes are displayed in Table 3. Wa can therefore conclude that the first constraint parameter τ_1 has a small influence on the metrics, both with and without NMS. Nevertheless, a higher performance is reached in the balanced case, *i.e.*, when $\tau_1 \to +\infty$.

H. Timing Analysis for SSD

As can be seen in Table 4, OT-based matches improve the epoch time (forward pass, compute the match cost, matching algorithm, and backward pass; in blue) for SSD with the Hungarian algorithm by almost 50%. The difference is smaller for DETR and variants as the models are proportionally heavier and the number of predictions smaller.

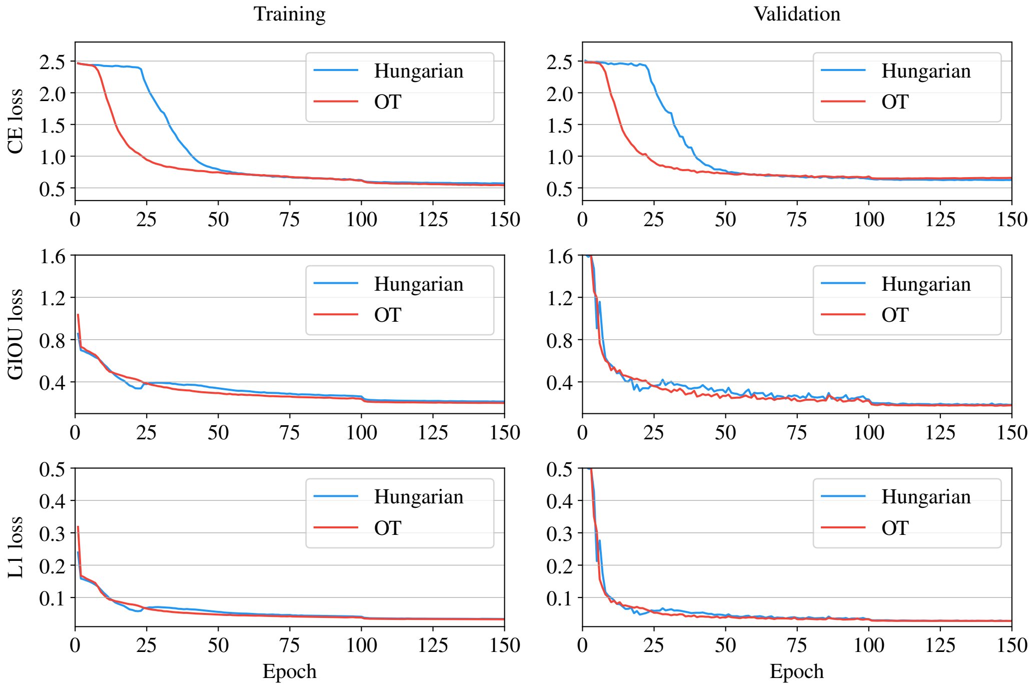


Figure 12. Training and validation unscaled loss curves for DETR on the Color Boxes dataset. The training loss is the average over the epoch.

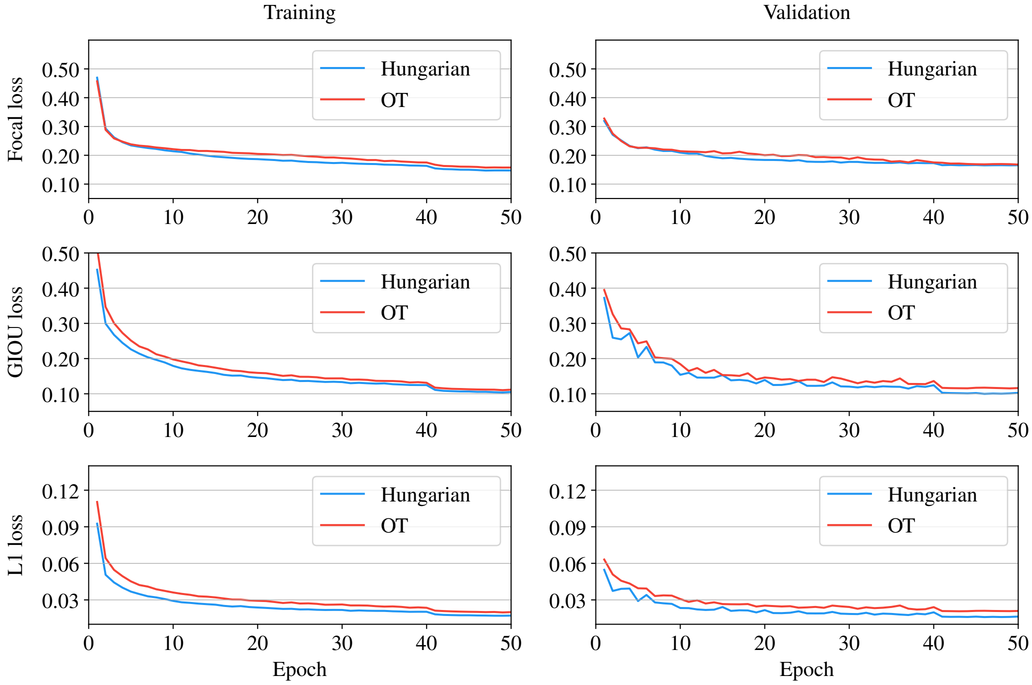
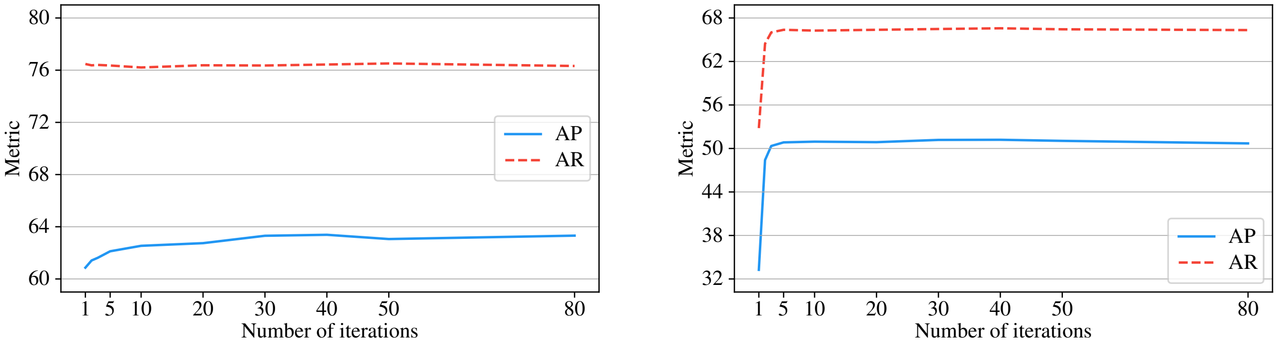


Figure 13. Training and validation unscaled loss curves for Deformable DETR on the Color Boxes dataset. The training loss is the average over the epoch.



(a) Deformable DETR with Balanced OT.

(b) SSD300 with Unbalanced OT ($\tau_2 = 0.01$).

Figure 14. Influence of the number of Sinkhorn iterations on the final metrics on the Color Boxes dataset.

Matching	τ_1	**with NMS**		**w/o NMS**	
		AP	**AR**	**AP**	**AR**
Unb. OT	0.01	47.2	62.0	41.9	71.1
Unb. OT	0.1	47.7	63.7	44.7	72.3
Unb. OT	1	47.7	64.0	44.8	72.7
Unb. OT	10	47.8	63.8	45.0	72.6
OT	(∞)	**48.1**	**64.3**	**45.2**	**73.0**

Table 3. Comparison of matching strategies on the Color Boxes dataset. SSD300 is evaluated both with and without NMS.

Epoch step	**OT**	**Unb. OT**	**Hung.**	**2-step**
Preprocessing	6.3 ms	*idem*	*idem*	*idem*
Forward pass	5.8 ms	*idem*	*idem*	*idem*
Anchor gen.	54.2 ms	*idem*	*idem*	*idem*
Match cost	4.2 ms	*idem*	*idem*	*idem*
Matching	1.1 ms	1.5 ms	18.3 ms	2.3 ms
Backward pass	8.2 ms	*idem*	*idem*	*idem*
Final losses	11.6 ms	11.6 ms	9.7 ms	9.7 ms

Table 4. Timing for each step in SSD300 on Color Boxes and a batch size of 16, computed with an Nvidia TITAN X GPU and Intel Core i7-4770K CPU @ 3.50GHz. Likewise the models we built upon, we used *Torchvision's* anchor generation implementation, which extensively relies on heavy loops and could drastically be improved (not the focus of our work). The final losses timings are partially due to the expensive hard-negative mining.

I. Color Boxes Dataset

This section provides a discussion of the Color Boxes synthetic dataset. It is split into 4,800 training and 960 validation images of 500 × 400 pixels. Images have a gray background. We uniformly randomly draw between 0 and 30 rectangles of 20 different colors, which define the category of the rectangle. The dimension of the rectangles vary from 12 to 80 pixels and are uniformly randomly rotated. They are placed such that the IoU between their bounding boxes is at most 0.25. A gaussian noise of mean 0 and standard deviation 0.05 is added to each pixel value independently. Sample images are drawn in Fig. 15.

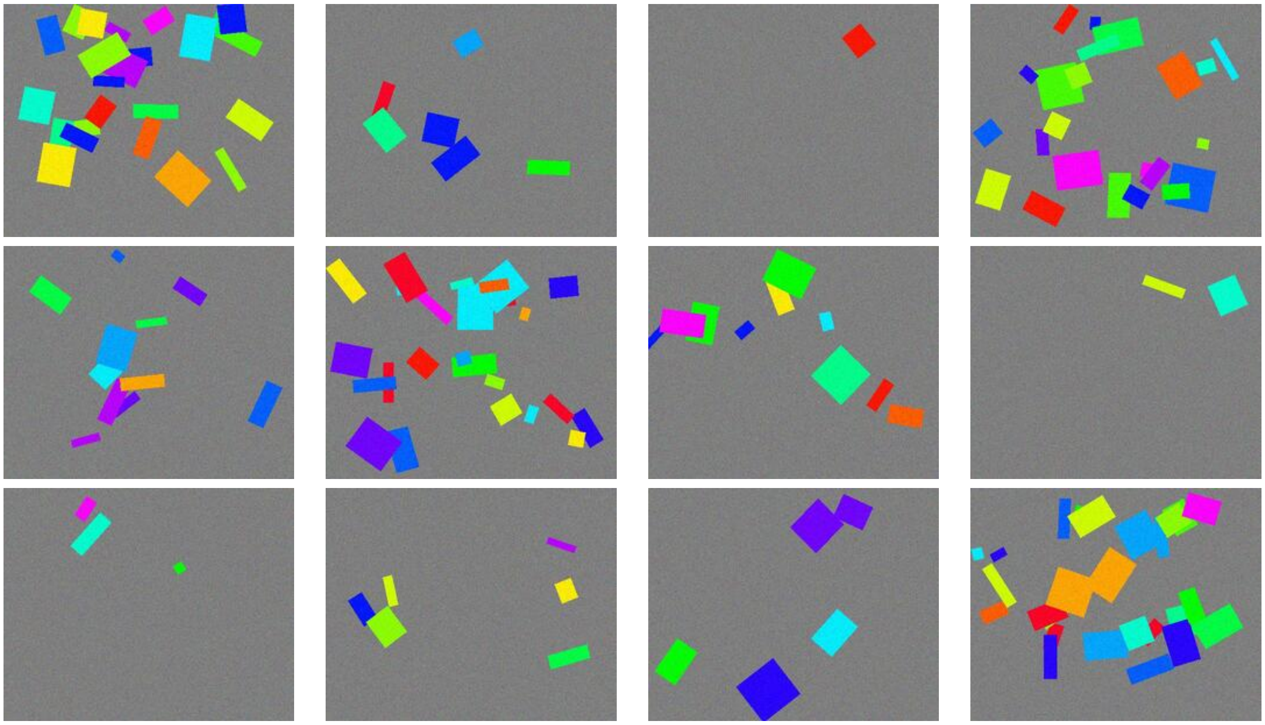


Figure 15. Sample images from the Color Boxes Dataset.

References

[1] Richard A. Brualdi. *Combinatorial Matrix Classes*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2006. 2

[2] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018. 4

[3] Lénaïc Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster wasserstein distance estimation with the sinkhorn divergence. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. 1

[4] Aude Genevay. *Entropy-Regularized Optimal Transport for Machine Learning*. Theses, PSL University, Mar. 2019. 1

[5] Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617. PMLR, 09–11 Apr 2018. 1

[6] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 1, 4, 9

[7] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 658–666, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society. 1

[8] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, pages 8922–8931, 2021. 4, 5