Favoring One Among Equals - Not a Good Idea: Many-to-one Matching for Robust Transformer based Pedestrian Detection

K.N Ajay Shastry[1] K. Ravi Sri Teja[1] Aditya Nigam[2] Chetan Arora[1]
[1]Indian Institute of Technology, Delhi
[2]Indian Institute of Technology, Mandi

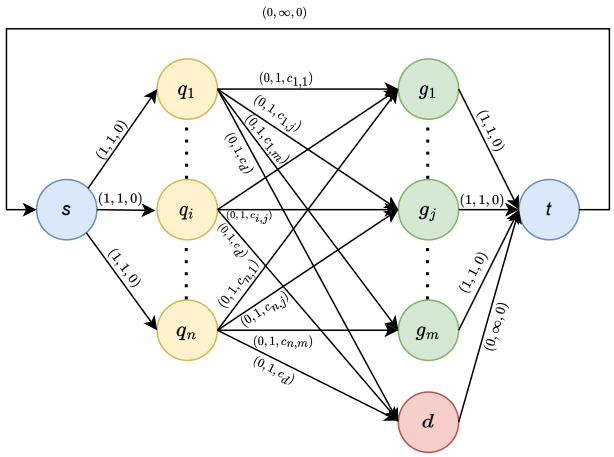


Figure 1. The constructed graph G to prove our claim. The tuple (x,y,z) on every edge denotes lower-capacity x, upper-capacity y and the cost z for every edge. The node colored in red indicates the dummy ground truth node. Matching a proposal with a dummy node indicates that there is no object corresponding to it.

Supplementary Material

1. Proof that one-to-one matching problem is a special case of min-cost-flow problem

To show that one-to-one matching is a special case of the min-cost-flow graph problem, we construct a graph G as discussed in Section-3.3.1, with the value of $k = 1$, where k indicates the upper capacity $u(g_j, t)\ \forall j \in \{1, 2, \ldots, m\}$. Fig. 1 shows the graph G. For our proof, we first show that the min-cost-flow algorithm on constructed graph G reduces to minimizing $\sum_j c_{\sigma(j),j}$ where $\sigma : \{1, 2, ..., m\} \rightarrow \{1, 2..., n\}$ is an injective function.

1.1. Proof that exactly one proposal node is matched to a ground truth node

From the graph G, we observe that the edges from ground truth g_j to the sink t, have a lower capacity $l(g_j, t) = 1$, and upper capacity $u(g_j, t) = 1\ \forall j \in \{1, 2, \ldots, m\}$. Therefore, in a valid flow obtained through a minimum cost flow algorithm, we obtain the flow at the edges (g_j,t) as follows:

$$f(g_j, t) = 1 \quad \forall j = \{1, 2...m\}. \tag{1}$$

Similarly edges from the source s to the proposals q_i have lower capacity $l(s, q_i) = 1$ and an upper capacity $u(s, q_i) = 1\ \forall i \in \{1, 2, \ldots, n\}$. Therefore, flow at the edges (s, q_i) must be:

$$f(s, q_i) = 1 \quad \forall i = \{1, 2...n\}. \tag{2}$$

Further, net flow at every ground truth node is:

$$f(g_j) = \sum_{\{b:(g_j,b)\in E\}} f(g_j, b) - \sum_{\{b:(b,g_j)\in E\}} f(b, g_j). \tag{3}$$

Substituting for all edges in G that are incident from and on a given ground truth g_j, we have:

$$f(g_j) = f(g_j, t) - \sum_i f(q_i, g_j). \tag{4}$$

We know from Eq. (1) that $f(g_j, t) = 1$. Therefore:

$$f(g_j) = 1 - \sum_i f(q_i, g_j). \tag{5}$$

Since net flow is zero at each node, hence, $f(g_j) = 0$, and

$$\sum_i f(q_i, g_j) = 1. \tag{6}$$

The integral flow theorem (Theorem-9.10 in [1]) guarantees that if $\forall (a, b) \in E$, $l(a, b) \in \mathbb{Z}$, and $u(a, b) \in \mathbb{Z}$, then $f(a, b) \in \mathbb{Z}$ after minimum cost flow computation. In our case, for all directed edges (q_i, g_j) in G, the lower capacity $l(q_i, g_j) = 0$, and the upper capacity $u(q_i, g_j) = 1$. Therefore, using integral flow theorem, we obtain that

$$f(q_i, g_j) \in \{0, 1\}\ \forall i \in \{1, \ldots, n\}, \forall j \in \{1, \ldots, m\}. \tag{7}$$

From Eq. (6), Eq. (7) we can infer that there exists exactly one value of q_i for any given g_j such that $f(q_i, g_j) = 1$ i.e., exactly one proposal node is matched to each ground truth node.

1.2. Uniqueness of proposal per ground truth

In this section, we show that there exists a unique proposal per ground truth. Let σ be the relation from $\{1, 2, \ldots, m\} \rightarrow \{1, 2, \ldots, n\}$ where m is the number of ground truths and n is the number of proposals defined as:

$$\sigma = \{(j, i) : f(q_i, g_j) = 1\}. \tag{8}$$

We show that σ is an injective function.

From the previous section, we obtain that there exists a single proposal per ground truth i.e., for any given g_j there exists q_i such that $f(q_i, g_j) = 1$. Therefore σ is a function from $\{1, 2, \ldots, m\}$ to $\{1, 2, \ldots, n\}$. To show that σ is injective, for any two given ground truth indices j_1, j_2 we need to show that if $\sigma(j_1) = \sigma(j_2)$, then $j_1 = j_2$. We prove this by contradiction.

Let us assume two distinct ground truth indices j_1, j_2 such that $\sigma(j_1) = \sigma(j_2) = i$ where i represents the index of a proposal. The net flow at q_i is:

$$f(q_i) = \sum_{\{b:(q_i,b)\in E\}} f(q_i, b) - \sum_{\{b:(b,q_i)\in E\}} f(b, q_i). \tag{9}$$

Substituting for all edges in G that are incident from and on a given proposal q_i, we have:

$$f(q_i) = \sum_j f(q_i, g_j) - f(s, q_i). \tag{10}$$

We know from Eq. (2) that $f(s, q_i) = 1$ and therefore we have

$$f(q_i) = \sum_j f(q_i, g_j) - 1. \tag{11}$$

Since, net flow is zero at each node, hence $f(q_i) = 0$, and:

$$\sum_j f(q_i, g_j) = 1. \tag{12}$$

On expanding Eq. (12), we obtain:

$$f(q_i, g_{j_1}) + f(q_i, g_{j_2}) + \sum_{j\neq j_1, j_2} f(q_i, g_j) = 1. \tag{13}$$

Since $\sigma(j_1) = \sigma(j_2) = i$, we have:

$$f(q_i, g_{j_1}) = f(q_i, g_{j_2}) = 1. \tag{14}$$

Therefore, we obtain that:

$$1 + 1 + \sum_{j\neq j_1, j_2} f(q_i, g_j) = 1 \tag{15}$$

$$\sum_{j\neq j_1, j_2} f(q_i, g_j) = -1 \tag{16}$$

We know from Eq. (7), that $f(q_i, g_j)$ takes values from $\{0, 1\}$. Therefore $\sum_{j\neq j_1, j_2} f(q_i, g_j) \geq 0$ for any given proposal q_i, which is in contradiction to Eq. (16). This implies that there cannot exist two distinct ground truth indices j_1, j_2 such that $\sigma(j_1) = \sigma(j_2) = i$. This also concludes that j_1 must be equal to j_2, and hence σ must be an injective function, and that there exists a unique proposal for every ground truth.

1.3. Matching Cost Optimization

The net flow at source node s can be written as:

$$f(s) = \sum_{\{b:(s,b)\in E\}} f(s, b) - \sum_{\{b:(b,s)\in E\}} f(b, s) \tag{17}$$

Substituting for all edges in G that are incident from and on s, we have:

$$f(s) = \sum_i f(s, q_i) - f(t, s) \tag{18}$$

We know from Eq. (2) that $f(s, q_i) = 1$. Therefore:

$$f(s) = \sum_i 1 - f(t, s) = n - f(t, s) \tag{19}$$

Since net flow at each node is zero. Therefore, $f(s) = 0$, and:

$$f(t, s) = n \tag{20}$$

Similarly, for the sink node t, the net flow is:

$$f(t) = \sum_{\{b:(t,b)\in E\}} f(t, b) - \sum_{\{b:(b,t)\in E\}} f(b, t). \tag{21}$$

Substituting for all edges in G that are incident from and on t, we have:

$$f(t) = f(t, s) - \left(\sum_j f(g_j, t) + f(d, t) \right). \tag{22}$$

Here d denotes the dummy ground truth node. We know from Eq. (20) that $f(t, s) = n$. Therefore:

$$f(t) = n - \sum_j f(g_j, t) - f(d, t). \tag{23}$$

We know from Eq. (1) that $f(g_j, t) = 1$. Therefore,

$$f(t) = n - \sum_j 1 - f(d, t) \tag{24}$$

Since the number of ground truths is m, we obtain

$$f(t) = n - m - f(d, t) \tag{25}$$

Since net flow is zero at each node, $f(t) = 0$, and:

$$f(d, t) = n - m \tag{26}$$

Net flow at dummy ground truth d can be written as:

$$f(d) = \sum_{\{b:(d,b)\in E\}} f(d,b) - \sum_{\{b:(b,d)\in E\}} f(b,d) \quad (27)$$

Substituting for all edges in G that are incident from and on d:

$$f(d) = f(d,t) - \sum_i f(q_i,d) \quad (28)$$

We know from Eq. (26) that $f(d,t) = n - m$. Therefore:

$$f(d) = (n-m) - \sum_i f(q_i,d) \quad (29)$$

Since net flow is zero at each node, $f(d) = 0$, and:

$$\sum_i f(q_i,d) = n - m. \quad (30)$$

Total flow cost for the graph G can be computed as:

$$C = \sum_{(a,b)\in E} c(a,b).f(a,b) \quad (31)$$

Substituting for all edges in G, we obtain:

$$\begin{aligned} C = &\sum_i \sum_j c(q_i,g_j)f(q_i,g_j) + \sum_i c(s,q_i)f(s,q_i) \\ &+ \sum_i c(q_i,d)f(q_i,d) + \sum_j c(g_j,t)f(g_j,t) \\ &+ c(d,t)f(d,t) + c(t,s)f(t,s) \end{aligned} \quad (32)$$

Since the cost of all incoming and outgoing edges from the source s and sink t is zero, we can rewrite the equation as:

$$C = \sum_i \sum_j c(q_i,g_j)f(q_i,g_j) + \sum_i c(q_i,d)f(q_i,d) \quad (33)$$

Since the cost of all edges from proposals q_i to dummy ground truth d have the same cost c_d, we can rewrite the equation as:

$$C = \sum_i \sum_j c(q_i,g_j)f(q_i,g_j) + \sum_i c_d.f(q_i,d) \quad (34)$$

By construction, we have $c(q_i,g_j) = c_{i,j}$. Therefore:

$$C = \sum_i \sum_j c_{i,j}f(q_i,g_j) + c_d \sum_i f(q_i,d) \quad (35)$$

We know from Eq. (30) that $\sum_i f(q_i,d) = n - m$. Therefore:

$$C = \sum_i \sum_j c_{i,j}f(q_i,g_j) + c_d(n-m) \quad (36)$$

$$\begin{aligned} &= \sum_{i=\sigma(j)} \sum_j c_{i,j}f(q_i,g_j) + \sum_{i\neq\sigma(j)} \sum_j c_{i,j}f(q_i,g_j) \\ &+ c_d(n-m) \end{aligned} \quad (37)$$

From the definition of σ in Eq. (8), we know that $f(q_i,g_j) = 1$ if and only if $\sigma(j) = i$ and $f(q_i,g_j) = 0$ in all other cases. Therefore we obtain the total flow cost in the graph G as:

$$C = \sum_j c_{\sigma(j),j} + c_d(n-m). \quad (38)$$

Hence when we minimize C, we obtain

$$\begin{aligned} \min C &= \min \left(\sum_j c_{\sigma(j),j} + c_d(n-m) \right) \\ &= \min \left(\sum_j c_{\sigma(j),j} \right) + c_d(n-m) \end{aligned} \quad (39)$$

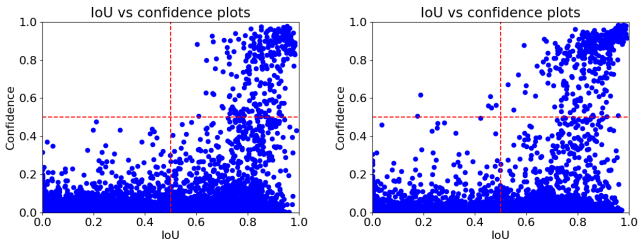
Since c_d, n, m are all constants, in order to minimize C, it is sufficient to minimize $\sum_j c_{\sigma(j),j}$. Hence our min-cost-flow formulation computes an injective function $\sigma : \{1,2,...,m\} \rightarrow \{1,2...,n\}$, which minimizes $\sum_j c_{\sigma(j),j}$. Thus, we prove that the one-to-one matching problem is a special case of a min-cost-flow problem.

2. Ablation study to understand the impact of proposed loss

To determine the effect of the proposed classification loss on our model, we conduct the following experiment in which 100 random images from the ECP [2] validation dataset are processed with DINO [6] and our proposed model. Then, we plot the graph of Confidence versus IoU for both. Confidence represents the prediction confidence for a specific object, and IoU represents the intersection over the union of the prediction and the actual ground truth. Fig. 2 depicts a graphical comparison of the two models' outputs. Fig. 2b reveals that with the introduction of proposed loss, the number of points in the upper right quadrant of the graph increases by 21% in comparison to that of Fig. 2a, indicating that with the introduction of this loss, the number of predictions with a high classification score and a high IoU with the ground truth has increased.

3. Ablation study to determine cost of matching with the dummy ground truth

To better understand the cost of matching with the dummy ground truth c_d, we performed the following experiment where we simulated the many-to-one matching by repeating the ground truth and then applied Hungarian matching at the RPN layer similar to [3]. For one of the intermediate training epochs, we plotted a histogram of the cost values for those matches with an IoU of less than 0.3 with the ground truth. Fig. 3 depicts the resulting histogram. We observe that the mean of this histogram is 2.9.



(a) DINO (b) Ours

Figure 2. The plot visualizes the confidence and IoU's distribution of matched samples in DINO and Ours. It can be observed that the proposed model has more predictions in the top right quadrant.

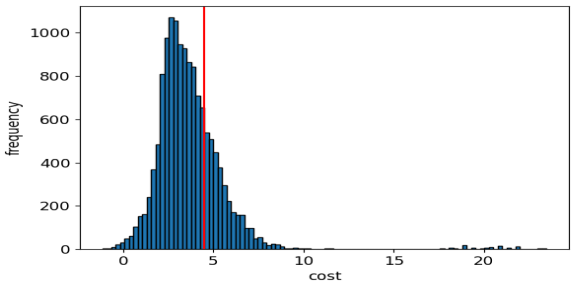


Figure 3. The plot shows the histogram of matching costs for those matches whose IoU between the proposal and the ground truth is less than 0.3

From our proposed flow-based matching strategy, a dummy ground truth vertex was introduced with an edge from all the proposals with an edge cost of c_d. We obtained our best results when c_d=4.5. Consequently, with c_d=4.5, we can infer from the histogram that it eliminates approximately 20% of those matches whose IoU with the ground truth is less than 0.3, previously produced by the many-to-one matching by repeating ground truth.

4. Comparision of the performance on COCO dataset

To demonstrate our model's performance on general object detection tasks, we trained it on the MS COCO Dataset [4]. Our model was trained from scratch using DINO's four-scale configuration over 12 epochs. We achieved a mAP of 57.4, which is an improvement of 0.6 mAP compared to DINO's vanilla version which scored an mAP of 56.8 in the same setting.

Model	Backbone	Epoch	mAP
DINO [6]	SwinL [5]	12	56.8
Ours	SwinL [5]	12	**57.4**

Table 1. Performance comparison between Vanilla DINO [6] and our model on MS COCO [4] Dataset

References

[1] Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice-Hall, Inc., USA, 1993. 1

[2] Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu M Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1844–1861, 2019. 3

[3] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19702–19712, 2023. 3

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4

[6] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Harry Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations*, 2022. 3, 4