HOTNAS: Hierarchical Optimal Transport for Neural Architecture Search

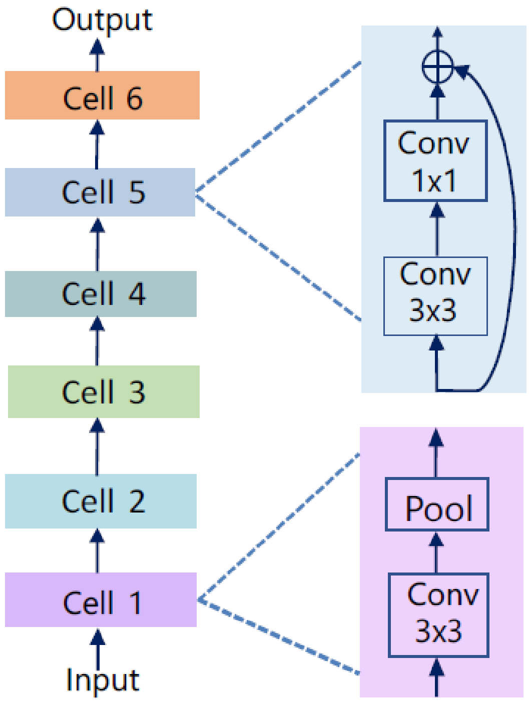
Poster: WED-AM-359

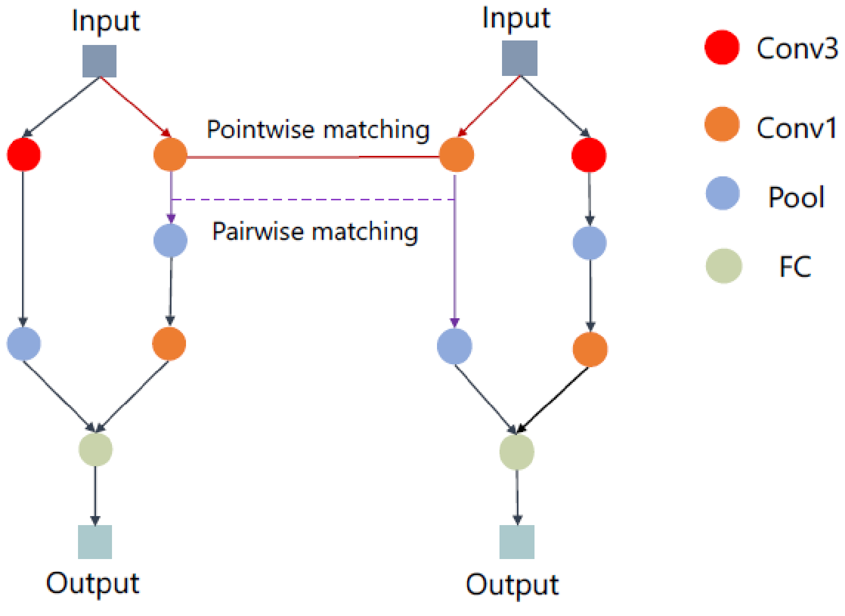
Jiechao Yang[1,2] Yong Liu[1,2] Hongteng Xu[1,2]

[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China

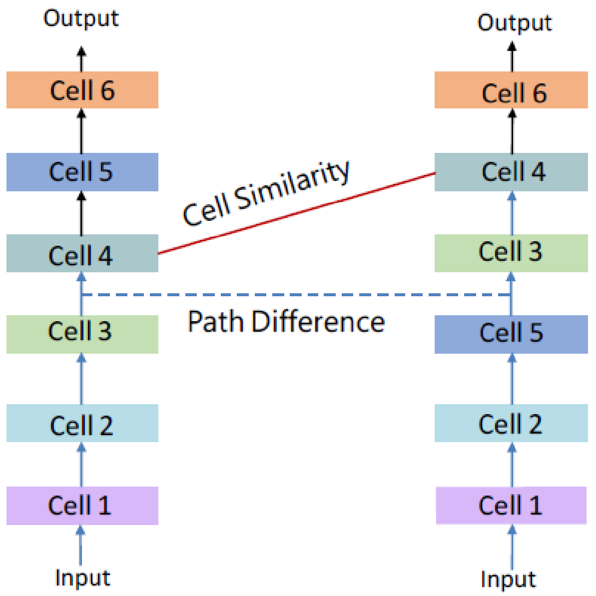
[2]Beijing Key Laboratory of Big Data Management and Analysis Methods

HOTNAS: Hierarchical Optimal Transport for Neural Architecture Search





(a) cell-level similarity



(b) network-level similary

A modular cell-based network with hierarchical structure

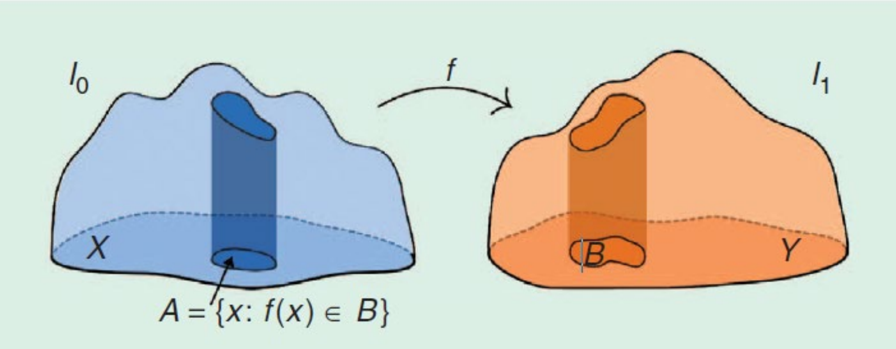
a **hierarchical optimal transportation meric** HOTNN， which jointly measures the similarity of cell internal architectures and the difference in macroarchitectures.

Background and related work

- The objective of NAS is to discover an optimal neural network architecture with the minimum validation loss

$$a^* = \underset{a \in \mathcal{A}}{\operatorname{argmin}} f(a),$$

- **Bayesian optimization (BO)** can **quickly** discover high-performing network architectures **with a limited number of samples**.



An example of OT [Kolouri et al. 2017]

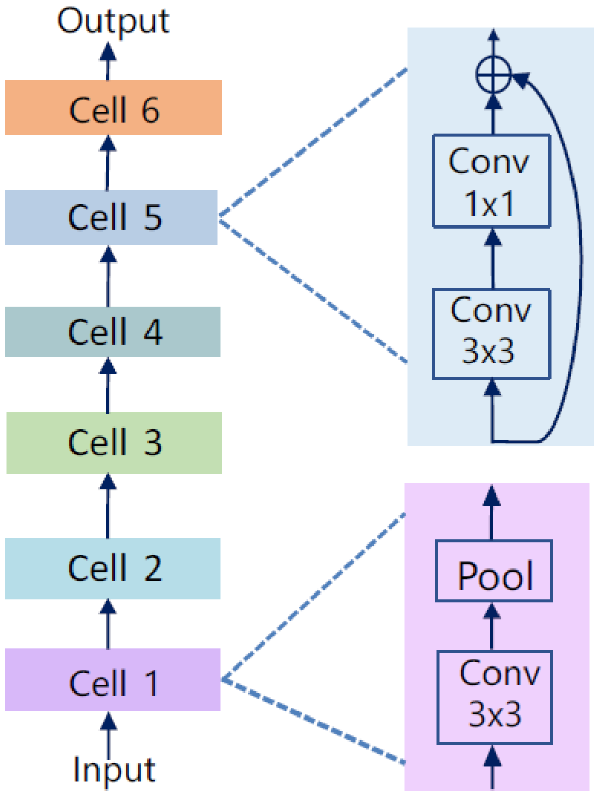
- The core of BO is **accurately measuring the similarity between different networks**.

  - Each network can be viewed as **a directed acyclic attributed graph.**
  - **Optimal Transport (OT)** can naturally handle the **graph-like architecture**.
  - NASBOT [Kandasamy et al. 2018] compute the minimum OT distance between networks as the similarity metric, but ignoring the similarity between cells.
  - TW [Nguyen et al. 2021] are limited to searching for a single cell architecture and ignore the similarity of macro-architectures.

Method

Network similarity metric： HOTNN

- measure the overall similarity between cell-based networks by leveraging its **hierarchical structure**
- organize the architecture into layers according to cells and learn the similarity within and between different layers.
- **cell-level similarity** computes **the OT distance between cells** in various networks by considering **the similarity of each node** and the **differences in the information flow costs between node pairs within each cell** in terms of **operational and structural information.**
- **Network-level similarity** calculates OT distance between networks by considering both **the cell-level similarity** and **the variation in the global position of each cell** within their respective networks



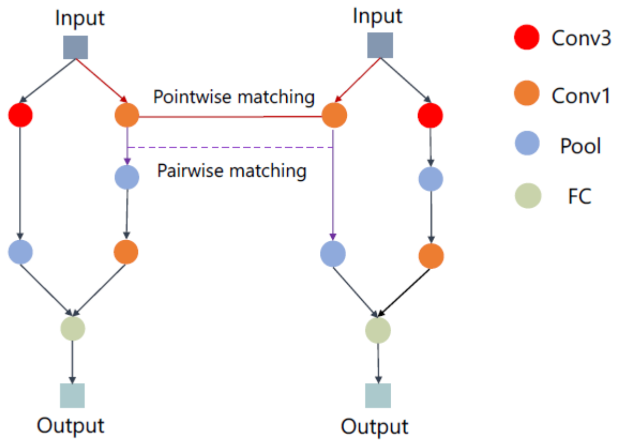
Example of a modular cell-based network

Cell-level similarity

- Two cell networks $\mathcal{G}_{B_1} = (\mathcal{L}^{B_1}, \mathcal{E}^{B_1}, \ell_o^{B_1}, \ell_s^{B_1})$, $\mathcal{G}_{B_2} = (\mathcal{L}^{B_2}, \mathcal{E}^{B_2}, \ell_o^{B_2}, \ell_s^{B_2})$

- Probability measure $\alpha = \sum_{i=1}^{n} p_i \delta_{(o_i^p, s_i^p)}$ $\beta = \sum_{j=1}^{m} q_j \delta_{(o_j^q, s_j^q)}$

- Transport matrix $\mathbf{T} \in U(\mathbf{p}, \mathbf{q}) = \{\mathbf{T} \in \mathbb{R}_+^{n\times m} \mid \mathbf{T}\mathbf{1}_m = \mathbf{p},\ \mathbf{T}^T\mathbf{1}_n = \mathbf{q}\}$,

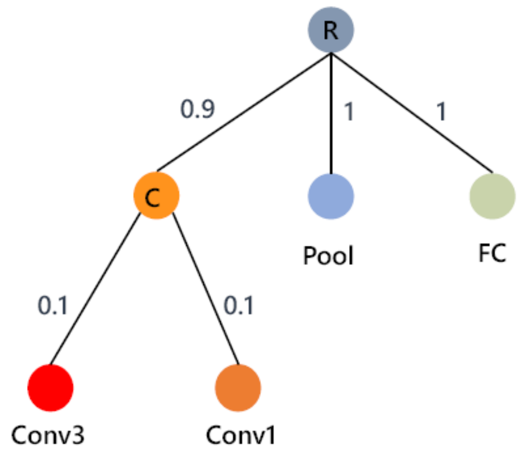


- **Pointwise matching:** considering differences **in the operational type and structural location information of each node.**

$$\min_{\mathbf{T}\in U(\mathbf{p},\mathbf{q})} \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{T}_{ij} C_{ij}^{pq}.$$

$$C_{ij}^{pq} = \varepsilon D(o_i^p, o_j^q) + (1-\varepsilon) H(s_i^p, s_j^q), \quad 1 \le i \le n, 1 \le j \le m,$$

$$H(s_i^p, s_j^q) = \frac{1}{6}(s_i^p - s_j^q)\mathbf{1}_6,$$



$D(o_i^p, o_j^q)$

	Conv3	Conv1	Pool	FC
Conv3	0	0.1	2	2
Conv1	0.1	0	2	2
Pool	2	2	0	2
FC	2	2	2	0

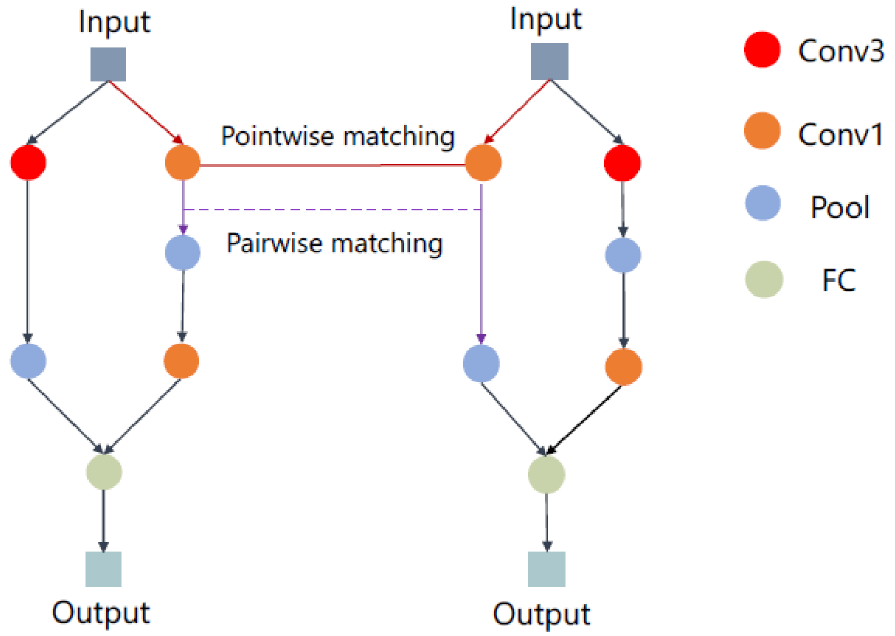
Cell-level similarity

- **Pairwise matching：** learn **the differences in the movement cost of various information flows between pairs of nodes within each cell network** in terms of **operational and structural information**.

$$\min_{\mathbf{T}\in U(\mathbf{p},\mathbf{q})} \sum_{i=1}^{n}\sum_{k=i+1}^{n}\sum_{j=1}^{m}\sum_{l=j+1}^{m} \mathbf{T}_{ij}\mathbf{T}_{kl}\left|C^p_{ik}-C^q_{jl}\right|.$$

$$C^p_{ik} = \varepsilon D(o^p_i, o^p_k) + (1-\varepsilon)H(s^p_i, s^p_k), 1 \le i < k \le n,$$

$$C^q_{jl} = \varepsilon D(o^q_j, o^q_l) + (1-\varepsilon)H(s^q_j, s^q_l), 1 \le j < l \le m.$$



- **iFGW metric**

$$\mathrm{iFGW}(\mathcal{G}_{B_1},\mathcal{G}_{B_2}) = \min_{\mathbf{T}\in U(\mathbf{p},\mathbf{q})} \sum_{i=1}^{n}\sum_{k=i+1}^{n}\sum_{j=1}^{m}\sum_{l=j+1}^{m} \lambda \mathbf{T}_{ij}C^p_{ij} + (1-\lambda)\mathbf{T}_{ij}\mathbf{T}_{kl}\left|C^p_{ik}-C^q_{jl}\right|,$$

Network-level similarity

- **Transport matching matrix**

$$\Gamma \in \mathrm{V}(\mathbf{f},\mathbf{g}) = \{\Gamma \in \mathbb{R}_{+}^{\mathrm{N}\times\mathrm{M}} \mid \Gamma\mathbf{1}_{\mathrm{M}} = \mathbf{f},\ \Gamma^{\mathrm{T}}\mathbf{1}_{\mathrm{N}} = \mathbf{g}\},$$

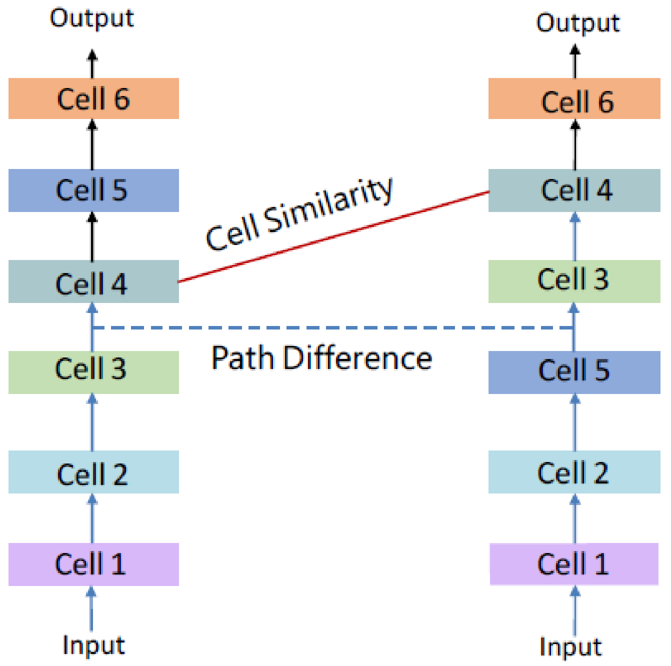
- **Cost matrix**: consider both the similarity between cells in two networks and the difference in the global position of each cell in their respective networks.

$$P(B_s^1, B_t^2) = \mid \delta^1(B_s^1)/\delta^1 - \delta^2(B_t^2)/\delta^2 \mid, \quad 1 \le s \le N, 1 \le t \le M,$$

$$\boldsymbol{S}_{st}^{12} = (1-\eta)\mathrm{iFGW}(B_s^1, B_t^2) + \eta P(B_s^1, B_t^2)$$

- **HOTNN metric**

$$\mathrm{HOTNN}(\boldsymbol{a}^1, \boldsymbol{a}^2) = \min_{\Gamma\in\mathrm{V}(\mathbf{f},\mathbf{g})} \sum_{s=1}^{N}\sum_{t=1}^{M} \Gamma_{\mathrm{st}} \boldsymbol{S}_{st}^{12}.$$



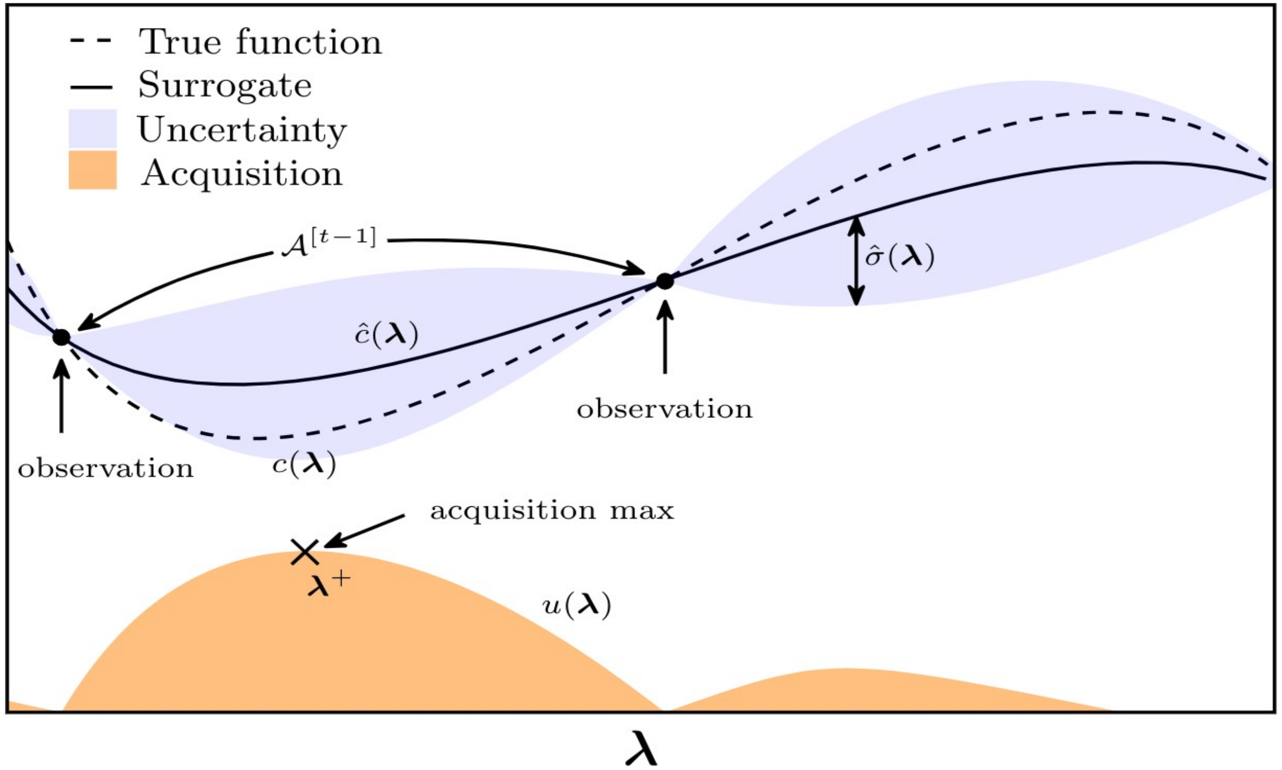
HOTNAS

Algorithm 1: Hierarchical Optimal Transport for Neural Architecture Search

Input: Total number of iterations N, initial datapoints $\mathcal{D}_0$, search space $\mathcal{A}$, The maxinum iterations T

Output: The best architecture $\boldsymbol{a}^*$

for $t = 0$ *to* $T - 1$ **do**
  Compute HOTNN metric between different architectures on the current observation set $\mathcal{D}_t = \mathcal{D}_0$;
  Embed the HOTNN metric to the kernel function of GP;
  Fit the GP on the current observation set $\mathcal{D}_t$;
  Construct the UCB acquisition function based on the predictive mean and variance (see Eq. (S7));
  Generate a pool of candidate architectures $\mathcal{P}_t$ by mutating the current best-performing architectures;
  Select the next promising architectures $\boldsymbol{a}_{\text{new}} = \arg\max_{\boldsymbol{a}\in\mathcal{P}_t} u_t(\boldsymbol{a})$;
  Evaluate $\boldsymbol{a}_{\text{new}}$ to obtain its validation loss y_{new};
  Update the observation set $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{\boldsymbol{a}_{\text{new}}, y_{\text{new}}\}$;
end
return the best-performing architecture $\boldsymbol{a}^* = \arg\min_{\boldsymbol{a}\in\mathcal{D}_T} f(\boldsymbol{a})$



Illustation of BO

Experiments

Table 1. Comparisons of the best-found valid loss and test loss on the TransNAS-Bench-101 benchmark and the DARTS benchmark.

Search Space	Tasks	Loss	Random Search	Evolutionary Search	BO-edit	NASBOT	HOTNAS
TransNAS-Bench-101	Autoencoding	Valid Error	28.09±0.18	28.19±0.24	28.55±0.28	29.25±0.25	**25.80±0.04**
		Test Error	26.65±0.18	26.74±0.24	27.14±0.28	27.82±0.25	**24.36±0.01**
	Object Classification	Valid Error	53.21±0.02	52.96±0.02	53.42±0.03	53.28±0.02	**52.69±0.00**
		Test Error	46.49±0.02	46.25±0.04	46.67±0.03	46.30±0.03	**45.90±0.02**
	Scene Classification	Valid Error	43.85±0.03	43.43±0.03	43.58±0.03	43.78±0.04	**43.19±0.01**
		Test Error	35.43±0.02	35.25±0.03	35.29±0.02	35.25±0.03	**35.00±0.01**
	Jigsaw	Valid Error	3.58±0.03	3.25±0.01	3.31±0.00	3.27±0.01	**3.17±0.01**
		Test Error	3.79±0.04	3.35±0.01	3.34±0.01	3.38±0.02	**3.29±0.01**
	Surface Normal	Valid Error	39.20±0.04	37.55±0.16	37.42±0.14	38.80±0.15	**36.65±0.20**
		Test Error	36.27±0.04	34.72±0.15	34.69±0.14	35.99±0.15	**33.91±0.19**
	Room Layout	Valid Error	59.98±0.04	59.83±0.03	59.95±0.06	60.19±0.05	**58.92±0.05**
		Test Error	53.94±0.06	55.54±0.10	54.02±0.03	54.72±0.09	**53.80±0.04**
	Semantic Segmentation	Valid Error	71.59±0.04	71.39±0.05	71.01±0.05	70.89±0.04	**70.51±0.01**
		Test Error	68.97±0.01	68.11±0.05	68.45±0.05	68.30±0.04	**67.98±0.03**
DARTS	CIFAR-10	Valid Error	5.90±0.07	5.50±0.09	5.42 ±0.14	5.73±0.07	**5.37±0.01**
		Test Error	3.28±0.09	2.87±0.04	2.72±0.07	2.93±0.12	**2.43±0.04**
	CIFAR-100	Test Error	21.47±0.08	19.75±0.13	20.62±0.12	19.95±0.17	**18.46±0.09**

Experiments

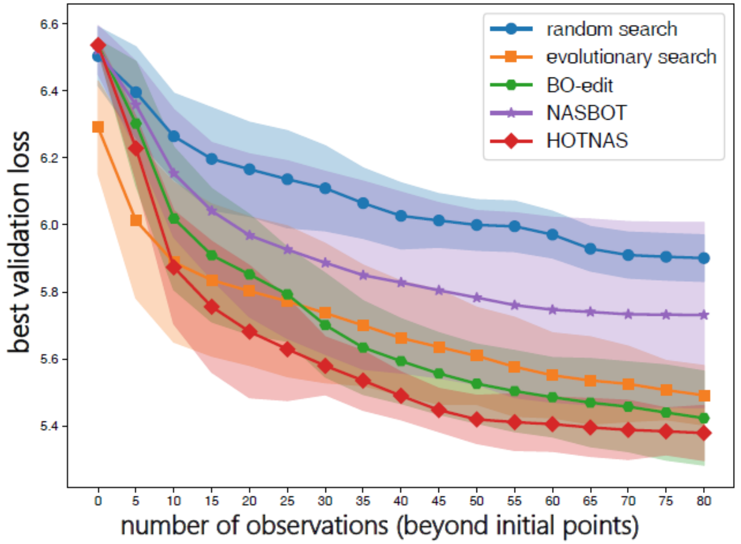
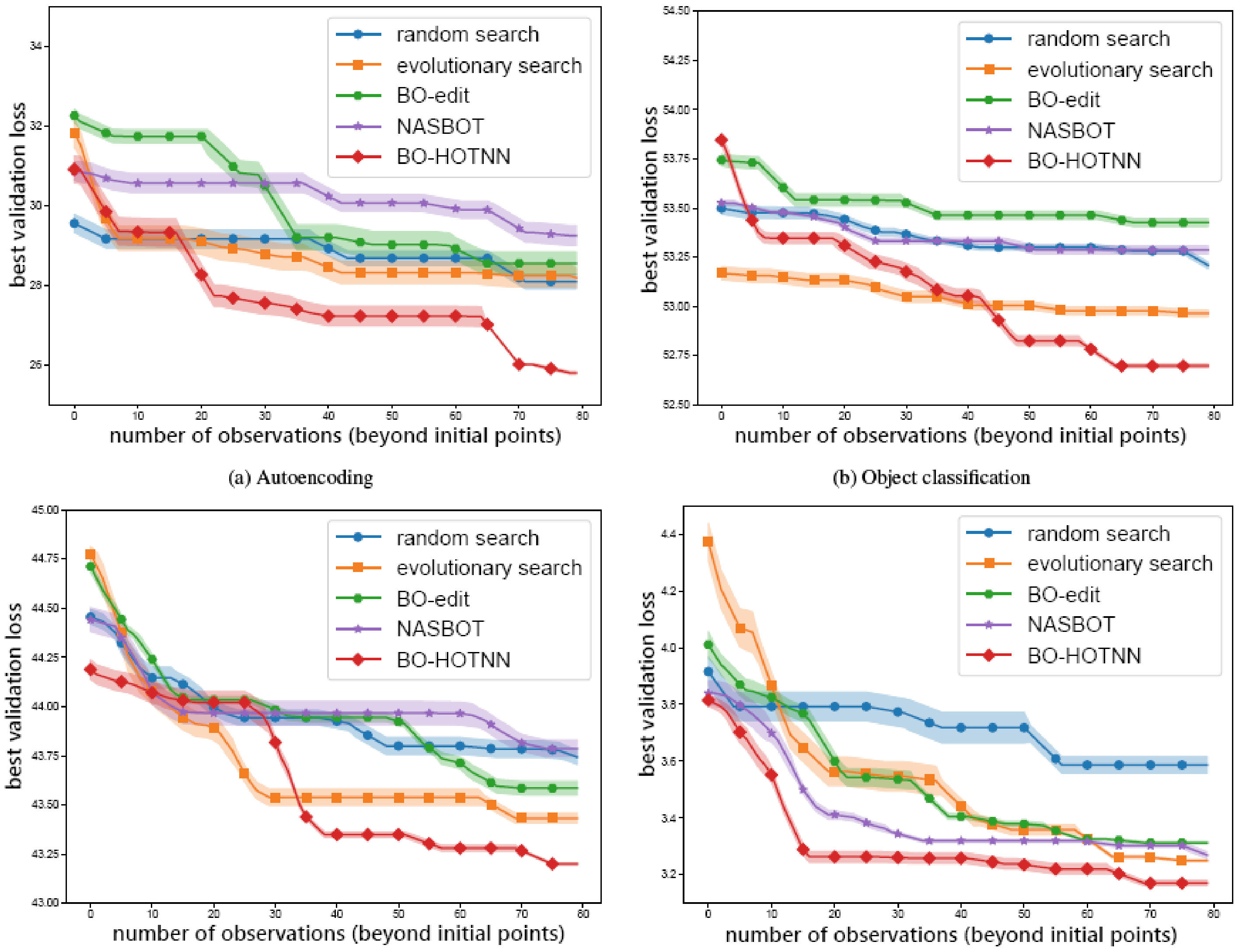


Figure 3. The best-found validation loss over the number of iterations (beyond initial points) of various NAS methods on the DARTS benchmark.



(a) Autoencoding

(b) Object classification

(c) Scene classification

(d) Jigsaw

Thanks!