THE CHIPS TO SYSTEMS CONFERENCE

61

SHAPING THE NEXT GENERATION OF ELECTRONICS

JUNE 23-27, 2024

MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA

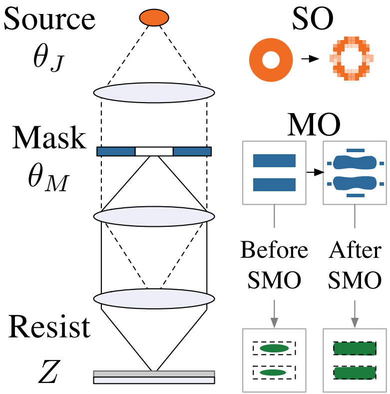
THE CHIPS TO SYSTEMS CONFERENCE
61
SHAPING THE NEXT GENERATION OF ELECTRONICS
JUNE 23-27, 2024
MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA

Efficient Bilevel Source Mask Optimization

Guojin Chen[1], Hongquan He[2], Peng Xu[1], Hao Geng[2], Bei Yu[1]

[1]Chinese University of Hong Kong
[2]ShanghaiTech University

The source mask optimization flow.



(a) The forward lithography and SMO process.

Forward imaging equation.

The scalar imaging equation under partially coherent illumination

$$I(x_1, y_1) = \iiint \iiint_{-\infty}^{\infty} J_C(x_0 - x_0', y_0 - y_0') \, M(x_0, y_0) \, M^*(x_0', y_0') \\ H(x_1 - x_0, y_1 - y_0) \, H^*(x_1 - x_0', y_1 - y_0') \, \mathrm{d}x_0 \, \mathrm{d}y_0 \, \mathrm{d}x_0' \mathrm{d}y_0', \tag{1}$$

Just for simplification:

$$\boldsymbol{I} \leftarrow \iiint \iiint \boldsymbol{J} \cdot \boldsymbol{M} \cdot \boldsymbol{M}^* \cdot \boldsymbol{H} \cdot \boldsymbol{H}^* \tag{2}$$

Abbe-based SMO: source and mask parameters: $\boldsymbol{\theta}_J$ and $\boldsymbol{\theta}_M$.

Abbe equation parameters

	Activation	Initialization
Mask $\boldsymbol{M}$	$\boldsymbol{M} = \sigma(\alpha_m \cdot \boldsymbol{\theta}_M)$	$\boldsymbol{\theta}_M(x, y) = m_0,$ if $M_0(x, y) = 1;$ else $-m_0.$
Source $\boldsymbol{J}$	$\boldsymbol{J} = \sigma(\alpha_j \cdot \boldsymbol{\theta}_J)$	$\boldsymbol{\theta}_J(f, g) = j_0,$ if $J_0(f, g) = 1;$ else $-j_0.$

Table: The activation and initialization for Abbe-imaging.

For projection lens:

$$H(f, g) = \begin{cases} 1, & \text{if } \sqrt{f^2 + g^2} \leq \frac{NA}{\lambda}, \\ 0, & \text{otherwise}, \end{cases} \quad (3)$$

For resist modeling:

$$\boldsymbol{Z} = \sigma(\beta \cdot (\boldsymbol{I} - I_{tr})), \quad (4)$$

Abbe-based SMO: loss function

L2 loss:

$$\mathcal{L}_2 = \|\boldsymbol{Z} - \boldsymbol{Z}_t\|^2. \tag{5}$$

PVB loss:

$$\mathcal{L}_{pvb} = \|\boldsymbol{Z}_{\max} - \boldsymbol{Z}_t\|^2 + \|\boldsymbol{Z}_{\min} - \boldsymbol{Z}_t\|^2. \tag{6}$$

SMO loss:

$$\mathcal{L}_{smo} = \mathcal{L}_{so} = \mathcal{L}_{mo} = \gamma\mathcal{L}_2 + \eta\mathcal{L}_{pvb}, \tag{7}$$

The SMO problem is thus defined as:

$$(\hat{\boldsymbol{\theta}}_J, \hat{\boldsymbol{\theta}}_M) = \underset{(\boldsymbol{\theta}_J, \boldsymbol{\theta}_M)}{\operatorname{argmin}} \mathcal{L}_{smo}(\boldsymbol{\theta}_J, \boldsymbol{\theta}_M), \tag{8}$$

Previous alternating minimization-based SMO

Algorithm Alternating Minimization-based SMO (`AM-SMO`)

---

for $k = 1, 2, 3, \ldots;$ **do** ▷ *Alternating SO & MO.*

  while not converged **do** ▷ *SO iterations.*

    $(\boldsymbol{\theta}_J)_k \leftarrow \text{argmin}_{\boldsymbol{\theta}_J} \mathcal{L}_{so}\left(\boldsymbol{\theta}_J, (\boldsymbol{\theta}_M)_{k-1}\right);$ ▷ *$\boldsymbol{\theta}_M$ is fixed.*

  while not converged **do** ▷ *MO iterations.*

    $(\boldsymbol{\theta}_M)_k \leftarrow \text{argmin}_{\boldsymbol{\theta}_M} \mathcal{L}_{mo}\left((\boldsymbol{\theta}_J)_k, \boldsymbol{\theta}_M\right);$ ▷ *$\boldsymbol{\theta}_J$ is fixed.*

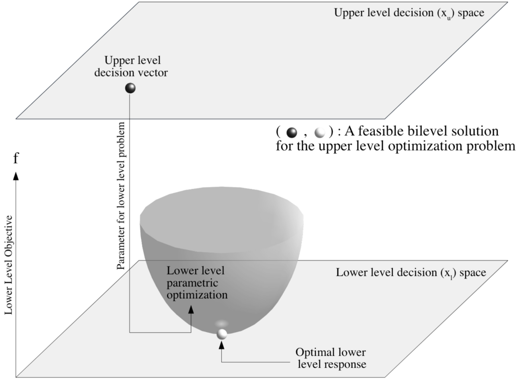
return $(\boldsymbol{\theta}_J)_k, (\boldsymbol{\theta}_M)_k$.

Drawbacks of traditional AM-SMO

However, `AM−SMO` has several notable **drawbacks**:

1. **Local-minima**: `AM−SMO` tends to converge to local minima due to its narrow focus on localized aspects of SO or MO, ignoring the global structure of the problem.
2. **Slow-convergence**: The convergence is often slow because the source and mask are highly interdependent. Adjusting $(\boldsymbol{\theta}_M)_k$ as per line 5 makes $(\boldsymbol{\theta}_J)_k$ suboptimal, requiring numerous iterations for stabilization.
3. **Lack of global perspective**: The absence of global gradient guidance complicates establishing effective early stopping criteria, often resulting in either prolonged optimization or suboptimal convergence.

Bilevel Optimization: Nested Scheme



- Leader takes a decision x
- The follower uses the leader's decision to take the best decision based on $f(x, \cdot)$
- The leader evaluates both x, y to evaluate $F(x, y)$.

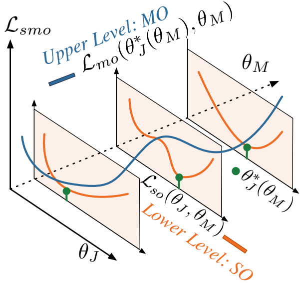
Reformulate into bilevel format.

$$(\hat{\boldsymbol{\theta}}_J, \hat{\boldsymbol{\theta}}_M) = \underset{(\boldsymbol{\theta}_J, \boldsymbol{\theta}_M)}{\text{argmin}}\ \mathcal{L}_{smo}(\boldsymbol{\theta}_J, \boldsymbol{\theta}_M), \quad (9)$$

$$\Downarrow \quad (10)$$

$$\min_{\boldsymbol{\theta}_M} \mathcal{L}_{mo}(\boldsymbol{\theta}_J^*(\boldsymbol{\theta}_M), \boldsymbol{\theta}_M), \quad \triangleright \textit{ Upper-Level: M}$$

$$\text{s.t. } \boldsymbol{\theta}_J^*(\boldsymbol{\theta}_M) = \text{argmin}_{\boldsymbol{\theta}_J} \mathcal{L}_{so}(\boldsymbol{\theta}_J, \boldsymbol{\theta}_M). \quad \triangleright \textit{ Lower-Level: S} \quad (11)$$



Solve the bilevel SMO : Hypergradient

$$\nabla_{\boldsymbol{\theta}_M} \mathcal{L}_{mo} = \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_M} + \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_J} \frac{\partial \boldsymbol{\theta}_J^*(\boldsymbol{\theta}_M)}{\partial \boldsymbol{\theta}_M}. \quad (12)$$

Direct gradient is easy to calculate: $\frac{\partial \mathcal{L}mo}{\partial \boldsymbol{\theta}_M}$ and $\frac{\partial \mathcal{L}mo}{\partial \boldsymbol{\theta}_J}$.

Challenges

1. Precise approximation of the SO optimal solution $\boldsymbol{\theta}_J^*(\boldsymbol{\theta}_M)$.
2. Differentiating the best-response Jacobian: $\frac{\partial \boldsymbol{\theta}_J^*(\boldsymbol{\theta}_M)}{\partial \boldsymbol{\theta}_M}$.

Three solutions to solve the bilevel SMO.

- `BiSMO-FD` : bilevel SMO with finite difference.
- `BiSMO-NMN`: bilevel SMO with Neumann series.
- `BiSMO-CG`: bilevel SMO with conjugate gradient.

BiSMO-FD: finite difference

To solve:

$$\nabla_{\boldsymbol{\theta}_M} \mathcal{L}_{mo} = \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_M} + \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_J} \frac{\partial \boldsymbol{\theta}_J^*(\boldsymbol{\theta}_M)}{\partial \boldsymbol{\theta}_M}. \tag{13}$$

BiSMO-FD

1. Approximate $\boldsymbol{\theta}_J^*(\boldsymbol{\theta}_M) = \boldsymbol{\theta}_J - \xi \nabla_{\boldsymbol{\theta}_J} \mathcal{L}_{so}$, (single step approximation)
2. obtaining $\frac{\partial \boldsymbol{\theta}_J^*(\boldsymbol{\theta}_M)}{\partial \boldsymbol{\theta}_M} = -\xi \frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_M \partial \boldsymbol{\theta}_J}$,
3. Hypergradient calculation:

$$\nabla_{\boldsymbol{\theta}_M} \mathcal{L}_{mo}^{\mathrm{FD}} = \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_M} - \xi \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_J} \frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_M \partial \boldsymbol{\theta}_J}, \tag{14}$$

IFT and BiSMO-NMN, BiSMO-CG

Implicit Function Theorem:

Consider $\boldsymbol{\theta}_J^*(\boldsymbol{\theta}_M)$, with **first-order optimally condition** $\frac{\partial \mathcal{L}_{so}(\boldsymbol{\theta}_J^*, \boldsymbol{\theta}_M)}{\partial \boldsymbol{\theta}_J} = 0$,

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\theta}_M}\left[\frac{\partial \mathcal{L}_{so}(\boldsymbol{\theta}_J^*(\boldsymbol{\theta}_M), \boldsymbol{\theta}_M)}{\partial \boldsymbol{\theta}_J}\right] = 0, \Rightarrow \frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_M \partial \boldsymbol{\theta}_J} + \frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_J \partial \boldsymbol{\theta}_J} \frac{\partial(\boldsymbol{\theta}_J^*(\boldsymbol{\theta}_M))}{\partial \boldsymbol{\theta}_M} = 0, \\
&\Rightarrow \text{best-response Jacobian: } \frac{\partial(\boldsymbol{\theta}_J^*(\boldsymbol{\theta}_M))}{\partial \boldsymbol{\theta}_M} = -\left[\frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_J \partial \boldsymbol{\theta}_J}\right]^{-1} \frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_M \partial \boldsymbol{\theta}_J}.
\end{aligned}
\tag{15}
$$

With Equation (12), we have hypergradient formulated by:

$$
\nabla_{\boldsymbol{\theta}_M} \mathcal{L}_{mo} = \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_M} - \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_J}\left[\frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_J \partial \boldsymbol{\theta}_J}\right]^{-1} \frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_M \partial \boldsymbol{\theta}_J}.
\tag{16}
$$

Challenges for IFT

$$\nabla_{\boldsymbol{\theta}_M}\mathcal{L}_{mo} = \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_M} - \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_J}\left[\frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_J \partial \boldsymbol{\theta}_J}\right]^{-1}\frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_M \partial \boldsymbol{\theta}_J}. \quad (17)$$

Challenges for IFT

- The inverse Hessian $\left[\frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_J \partial \boldsymbol{\theta}_J}\right]^{-1}$ in Equation (16) is hard to calculate.

Naive solution for ILT : `BiSMO-FD`

- In Equation (14), `BiSMO-FD` employs finite difference to naively approximate the inverse $\left[\frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_J \partial \boldsymbol{\theta}_J}\right]^{-1} = \xi \boldsymbol{\mathcal{I}}$, where $\boldsymbol{\mathcal{I}}$ denotes the identity matrix.

$$\nabla_{\boldsymbol{\theta}_M}\mathcal{L}_{mo}^{\mathrm{FD}} = \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_M} - \xi \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_J}\frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_M \partial \boldsymbol{\theta}_J}, \quad (18)$$

Better approx. of the inverse: `BiSMO-NMN` and `BiSMO-CG`

Do we have better approximation of the inverse?

More precise approximation of the inverse, IFT-based

- `BiSMO-NMN`: Neumann series
- `BiSMO-CG`: conjugate gradient

to reformulate the hypergradient.

Neumann series for inverse Hessian approximation

Lemma 2

Neumann series:
With a matrix $\boldsymbol{A}$ that $\|\boldsymbol{\mathcal{I}} - \boldsymbol{A}\| < 1$, we have:

$$\boldsymbol{A}^{-1} = \sum_{k=0}^{\infty} (\boldsymbol{\mathcal{I}} - \boldsymbol{A})^k. \tag{19}$$

Then we can approximate the inverse $\left[\frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_J \partial \boldsymbol{\theta}_J}\right]^{-1}$ and define the hypergradient as:

$$\nabla_{\boldsymbol{\theta}_M} \mathcal{L}_{mo} = \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_M} - \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_J} \sum_{k=0}^{\infty} \left[\boldsymbol{\mathcal{I}} - \frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_J \partial \boldsymbol{\theta}_J}\right]^k \frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_M \partial \boldsymbol{\theta}_J}. \tag{20}$$

Conjugate gradient: another way to approximate the inverse

- Specifically, $\frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_J} \left[\frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_J \partial \boldsymbol{\theta}_J}\right]^{-1}$ can be computed as the solution to the linear system $\left[\frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_J \partial \boldsymbol{\theta}_J}\right] \boldsymbol{w} = \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_J}$.
- The vector $\boldsymbol{w}$ can be obtained by solving the optimization problem:

$$\min_{\boldsymbol{w}} \boldsymbol{w}^\top \left[\frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_J \partial \boldsymbol{\theta}_J}\right] \boldsymbol{w} - \boldsymbol{w}^\top \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_J}. \tag{21}$$

- The conjugate gradient (`CG`) algorithm is well-suited for this task.

The hypergradient in Equation (16) for `BiSMO-CG` is then computed as:

$$\nabla_{\boldsymbol{\theta}_M} \tilde{\mathcal{L}}_{mo}^{\mathtt{CG}} = \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_M} - \left[\operatorname*{argmin}_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_J \partial \boldsymbol{\theta}_J} \boldsymbol{w} - \boldsymbol{w}^\top \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_J}\right)\right] \frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_M \partial \boldsymbol{\theta}_J}. \tag{22}$$

BiSMO overall flow

$$\text{Hypergradient: } \nabla_{\boldsymbol{\theta}_M}\mathcal{L}_{mo} = \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_M} + \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_J}\frac{\partial \boldsymbol{\theta}_J^*(\boldsymbol{\theta}_M)}{\partial \boldsymbol{\theta}_M}. \tag{23}$$

$$\star \text{ Finite Difference: } \nabla_{\boldsymbol{\theta}_M}\mathcal{L}_{mo}^{\mathrm{FD}} = \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_M} - \xi\frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_J}\frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_M \partial \boldsymbol{\theta}_J},$$

$$\S\text{IFT - Neumann Series: } \nabla_{\boldsymbol{\theta}_M}\tilde{\mathcal{L}}_{mo}^{\mathrm{NMN}} = \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_M} - \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_J}\sum_{k=0}^{K}\left[\boldsymbol{\mathcal{I}} - \frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_J \partial \boldsymbol{\theta}_J}\right]^k \frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_M \partial \boldsymbol{\theta}_J}.$$

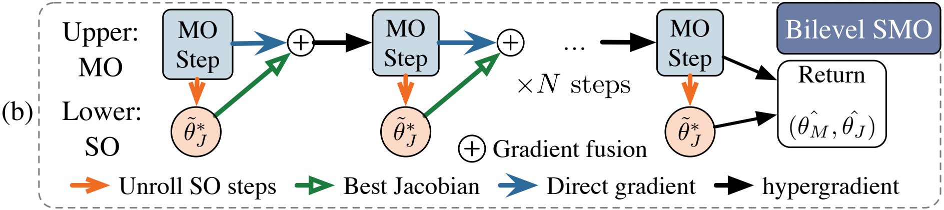
$$\S\text{IFT - Conjugate Gradient:} \nabla_{\boldsymbol{\theta}_M}\tilde{\mathcal{L}}_{mo}^{\mathrm{CG}} = \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_M} - \left[\underset{\boldsymbol{w}}{\operatorname{argmin}}\left(\boldsymbol{w}^\top \frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_J \partial \boldsymbol{\theta}_J}\boldsymbol{w} - \boldsymbol{w}^\top \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_J}\right)\right]\frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_M \partial \boldsymbol{\theta}_J}.$$

BiSMO overall flow

§ Finite Difference: $\nabla_{\boldsymbol{\theta}_M} \mathcal{L}_{mo}^{\mathrm{FD}} = \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_M} - \xi \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_J} \frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_M \partial \boldsymbol{\theta}_J},$

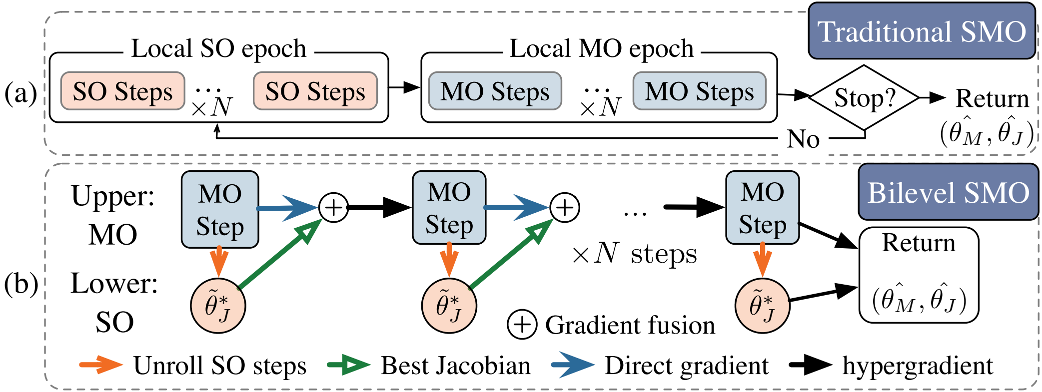
§ Neumann Series: $\nabla_{\boldsymbol{\theta}_M} \tilde{\mathcal{L}}_{mo}^{\mathrm{NMN}} = \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_M} - \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_J} \sum_{k=0}^{K} \left[\mathcal{I} - \frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_J \partial \boldsymbol{\theta}_J} \right]^k \frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_M \partial \boldsymbol{\theta}_J}.$

§ Conjugate Gradient: $\nabla_{\boldsymbol{\theta}_M} \tilde{\mathcal{L}}_{mo}^{\mathrm{CG}} = \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_M} - \left[\operatorname*{argmin}_{\boldsymbol{w}} \left(\boldsymbol{w}^\top \frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_J \partial \boldsymbol{\theta}_J} \boldsymbol{w} - \boldsymbol{w}^\top \frac{\partial \mathcal{L}_{mo}}{\partial \boldsymbol{\theta}_J} \right) \right] \frac{\partial^2 \mathcal{L}_{so}}{\partial \boldsymbol{\theta}_M \partial \boldsymbol{\theta}_J}.$



(b)Our BiSMO flow.

BiSMO vs. AM-SMO



Comparison of traditional AM-SMO and BiSMO.

Results visualization

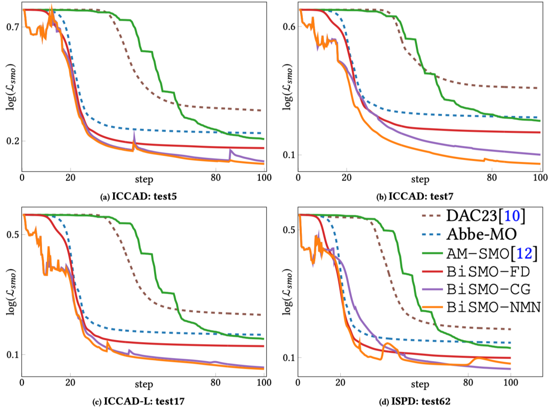
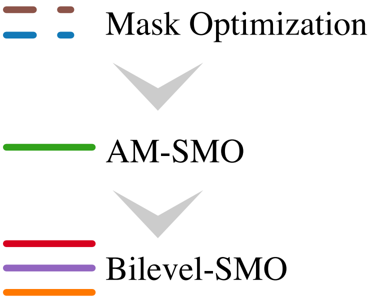


Figure 3: Loss comparison between different MO methods (dashed lines) and SMO methods (solid lines).



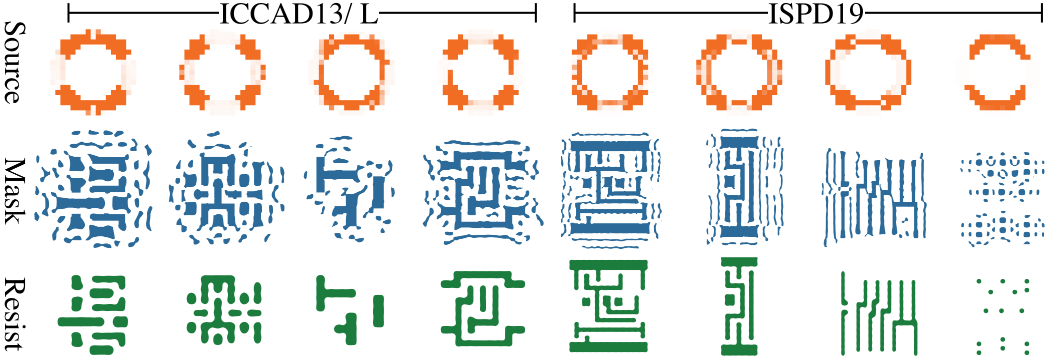
Comparison with SOTA

Table 3: Result comparison with SOTA.

Bench	MO						AM−SMO				BiSMO (Ours)					
	NILT [7]		DAC23-MILT [10]		Abbe-MO(Ours)		Abbe-Hopkins⋆ [13]		Abbe-Abbe [12]		BiSMO−FD		BiSMO−CG		BiSMO−NMN	
	L2	PVB	L2	PVB	L2	PVB	L2	PVB	L2	PVB	L2	PVB	L2	PVB	L2	PVB
ICCAD13	37515	50964	28362	40044	20419	29697	27299	37278	17539	23944	13828	17872	13603	16274	**13059**	**15839**
ICCAD-L	71570	108162	53143	87010	44478	66092	48879	77062	40455	58560	29779	42643	29762	40543	**28946**	**38706**
ISPD19	97891	119732	85234	105592	61374	93132	79634	97073	55588	84402	39959	64211	39488	61190	**38737**	**59832**
Average	68992	92953	55580	77549	42090	62974	51937	70471	37861	55635	27855	41576	27618	39336	**26914**	**38126**
Ratio	2.56	2.44	2.07	2.03	1.56	1.65	1.93	1.85	1.41	1.46	1.03	1.09	1.03	1.03	**1.00**	**1.00**

Abbe-Hopkins⋆ [13]: AM−SMO employs Abbe model for SO and Hopkins model for MO. L2 and PVB unit: nm^2.

Samples



Result samples from ICCAD13 and ISPD19 datasets.

THE CHIPS TO SYSTEMS CONFERENCE

SHAPING THE NEXT GENERATION OF ELECTRONICS

JUNE 23-27, 2024

MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA

Thanks!