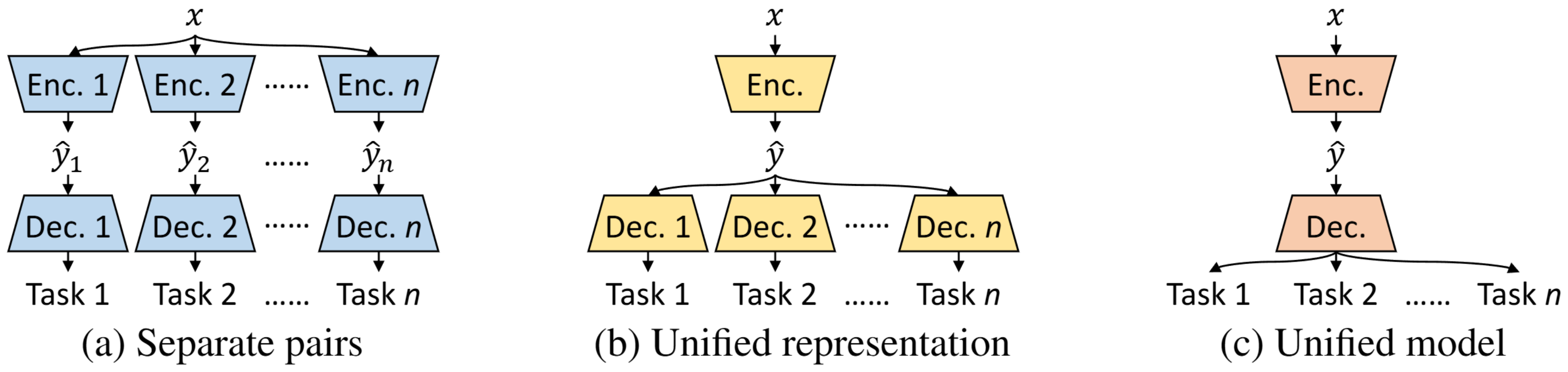
All-in-One Image Coding for Joint Human-Machine Vision with Multi-Path Aggregation

Xu Zhang, Peiyao Guo, Ming Lu, Zhan Ma

Vision Lab, Nanjing University

Background



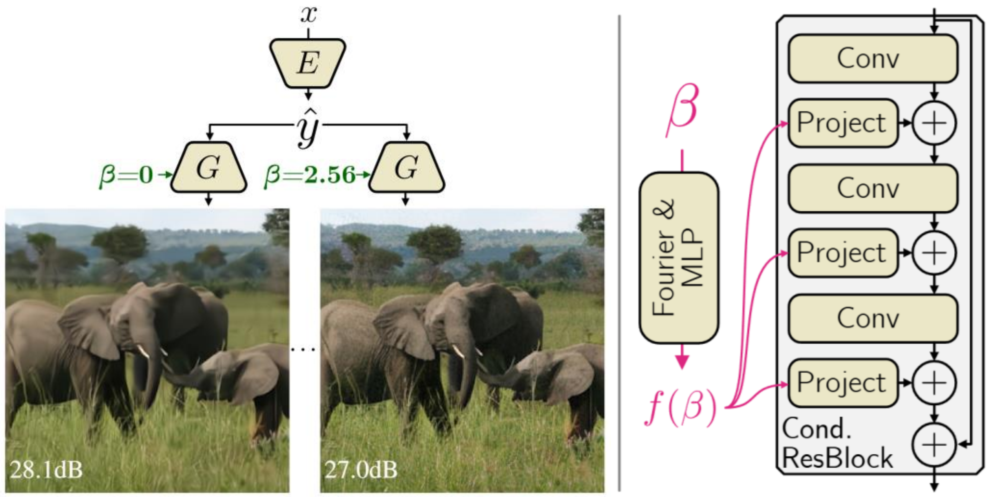
- 😊 Easy to optimize
- 😊 High accuracy
- ☹ Redundant bitstreams
- ☹ Redundant models

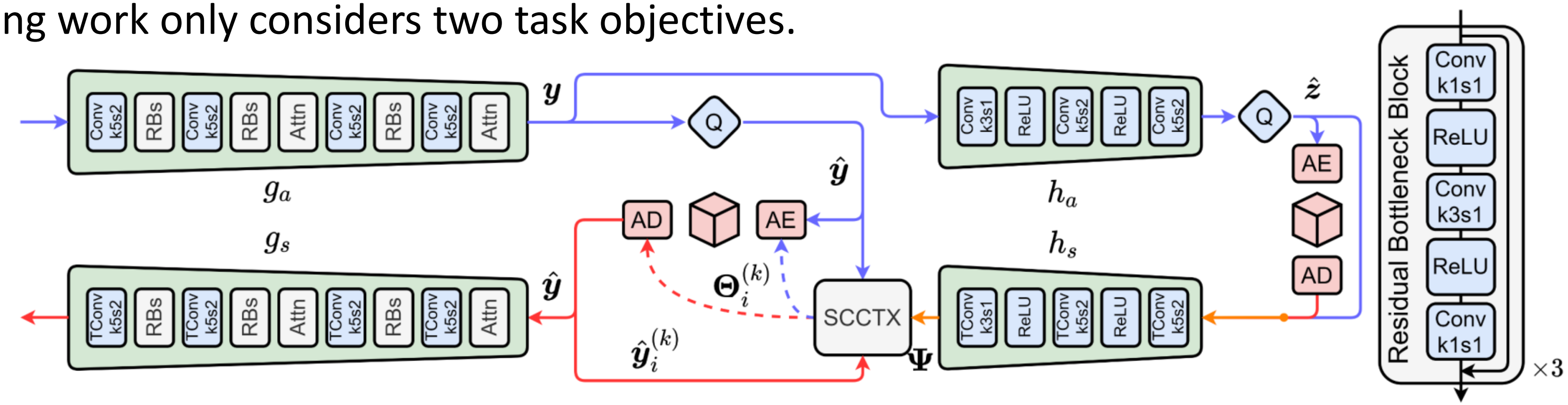
- 😊 Easy to optimize
- 😊 High accuracy
- 😊 Unified bitstream
- ☹ Redundant decoders

- 😊 Unified bitstream
- 😊 Unified model
- ☹ Hard to optimize
- ☹ Unstable performance

Background

- Existing unified approach: β-condition
- Pros:
  - Continuously Adjustable.
  - Supporting transitions between tasks.
  - Easily integrated into existing model frameworks.
- Cons:
  - Optimization is challenging, with variable hyperparams.
  - Balancing more tasks is difficult.
  - Existing work only considers two task objectives.

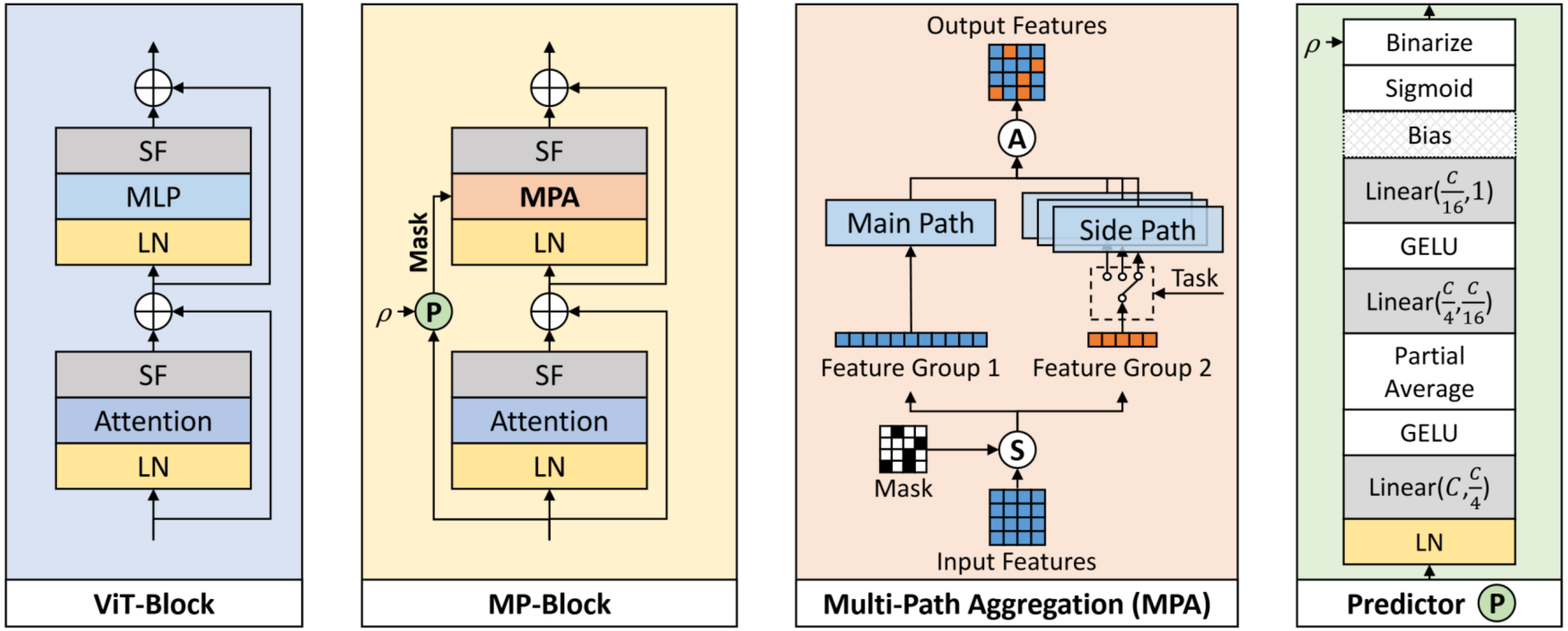




[1] D. He, et al. ELIC: Efficient Learned Image Compression with Unevenly Grouped Space-Channel Contextual Adaptive Coding. In CVPR, 2022.
[2] E. Agustsson, et al. Multi-Realism Image Compression With a Conditional Generator. In CVPR, 2023.

Our Approach

- Built upon existing models (TinyLIC is used in our implementation).
- Minimal additional components (only MLPs), reducing storage and computation overhead.
- Simple and effective optimization strategy, beneficial for task expansion.
- Supports smooth transitions between tasks.



Training Strategy

- Stage 1:
  - Optimize the base model with the Main Path.
  - The optimization objectives are Eqs. (6) and (7).
- Stage 2:
  - Optimize only the Side Path and Predictor.
  - The optimization objective is Eq. (8).
  - For MSE optimization, the task loss only includes MSE Loss.
  - For vision task optimization, the task loss is defined as Eq. (9).

$$\mathcal{L}_{\text{ratio}} = \frac{1}{S}\sum_{s=1}^{S}\left(\rho_{\text{enc}} - \frac{1}{H^{(s)}W^{(s)}}\sum_{h=1}^{H^{(s)}}\sum_{w=1}^{W^{(s)}} \boldsymbol{M}^{(s)}(h,w)\right)^2, \tag{4}$$

$$\mathcal{L}_G = \mathbb{E}_{\hat{\boldsymbol{y}}\sim p_{\boldsymbol{y}}}[-\log(D(\hat{\boldsymbol{y}}, G(\hat{\boldsymbol{y}})))], \tag{5}$$

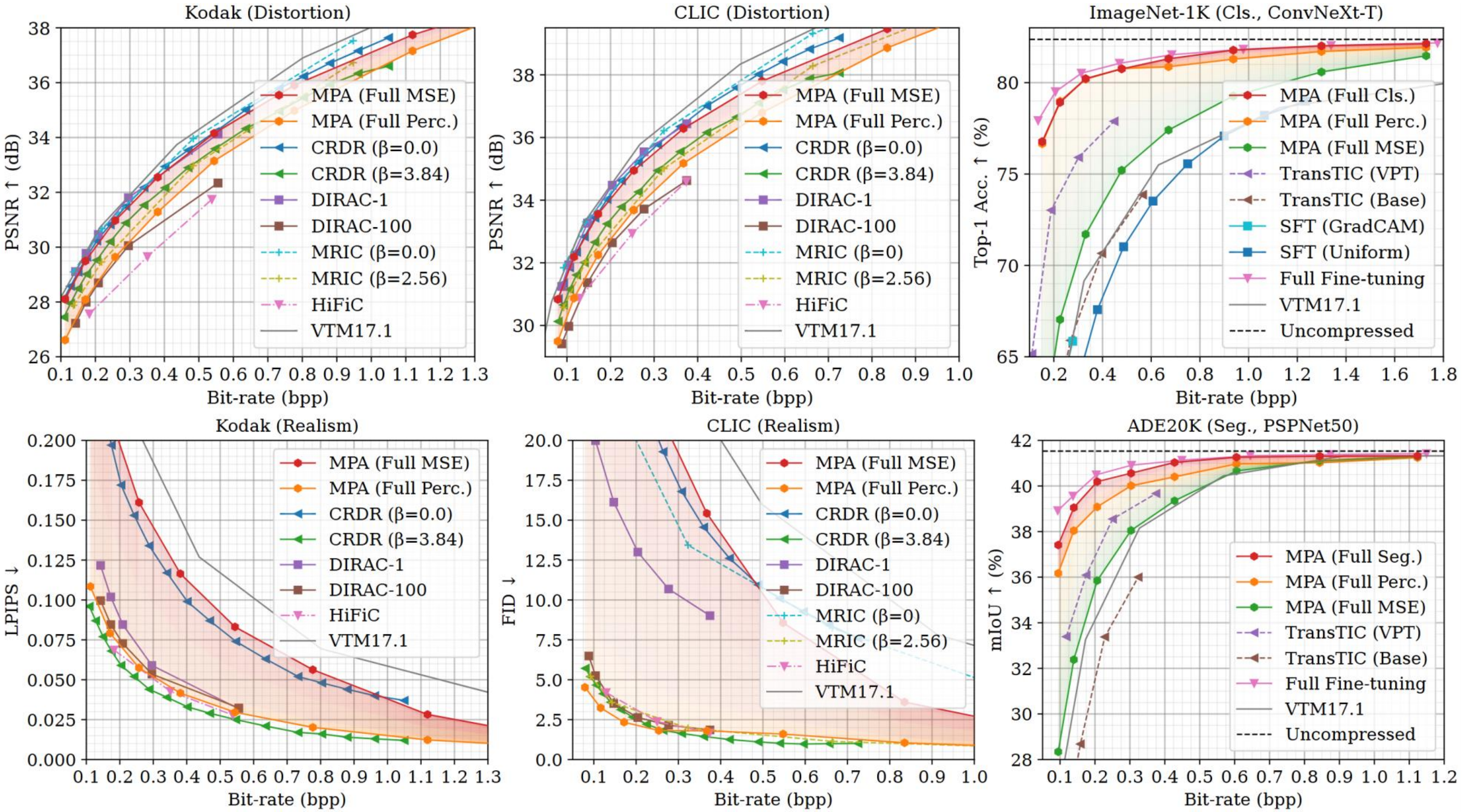
$$\mathcal{L}_D = \mathbb{E}_{\hat{\boldsymbol{y}}\sim p_{\boldsymbol{y}}}[-\log(1 - D(\hat{\boldsymbol{y}}, G(\hat{\boldsymbol{y}})))] + \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}}[-\log D(E(\boldsymbol{x}), \boldsymbol{x})], \tag{6}$$

$$\mathcal{L}_{EGP} = \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}}[\lambda_r^{(q)} r(\hat{\boldsymbol{y}}) + d(\boldsymbol{x}, \hat{\boldsymbol{x}})] + \lambda_G \mathcal{L}_G + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{ratio}}\mathcal{L}_{\text{ratio}}, \tag{7}$$

$$(\phi^*_{\text{side}}, \phi^*_{\text{pred}}) = \underset{\phi_{\text{side}},\phi_{\text{pred}}}{\arg\min}\, \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}}[\lambda_r^{(q)} r(\hat{\boldsymbol{y}})] + \lambda_{\text{task}}\mathcal{L}_{\text{task}} + \lambda_{\text{ratio}}\mathcal{L}_{\text{ratio}}, \tag{8}$$

$$\mathcal{L}_{\text{task}} = \text{CrossEntropy}(\text{ClsModel}(\text{Norm}(\hat{\boldsymbol{x}})), GT) + d(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}}. \tag{9}$$

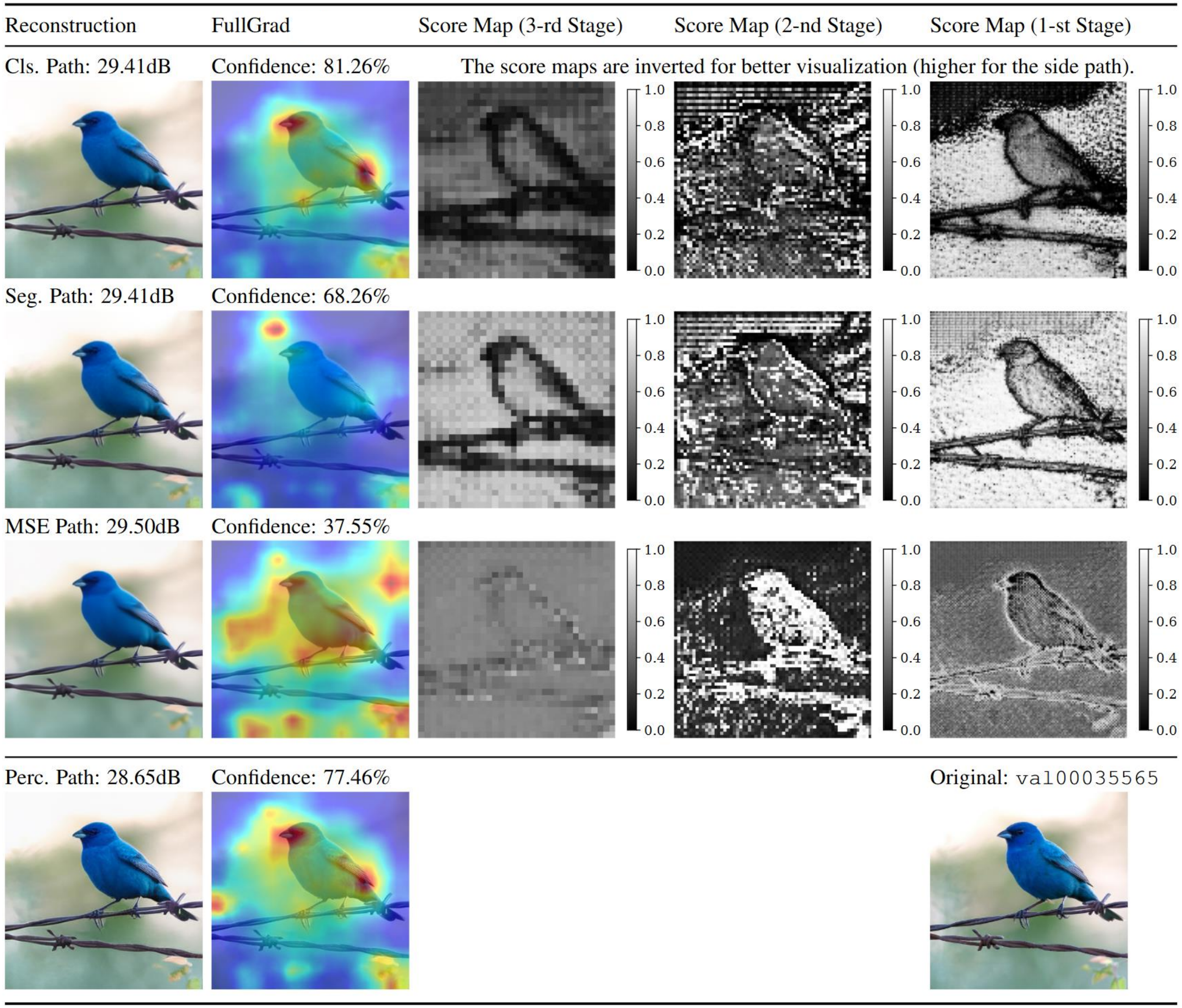
Performance



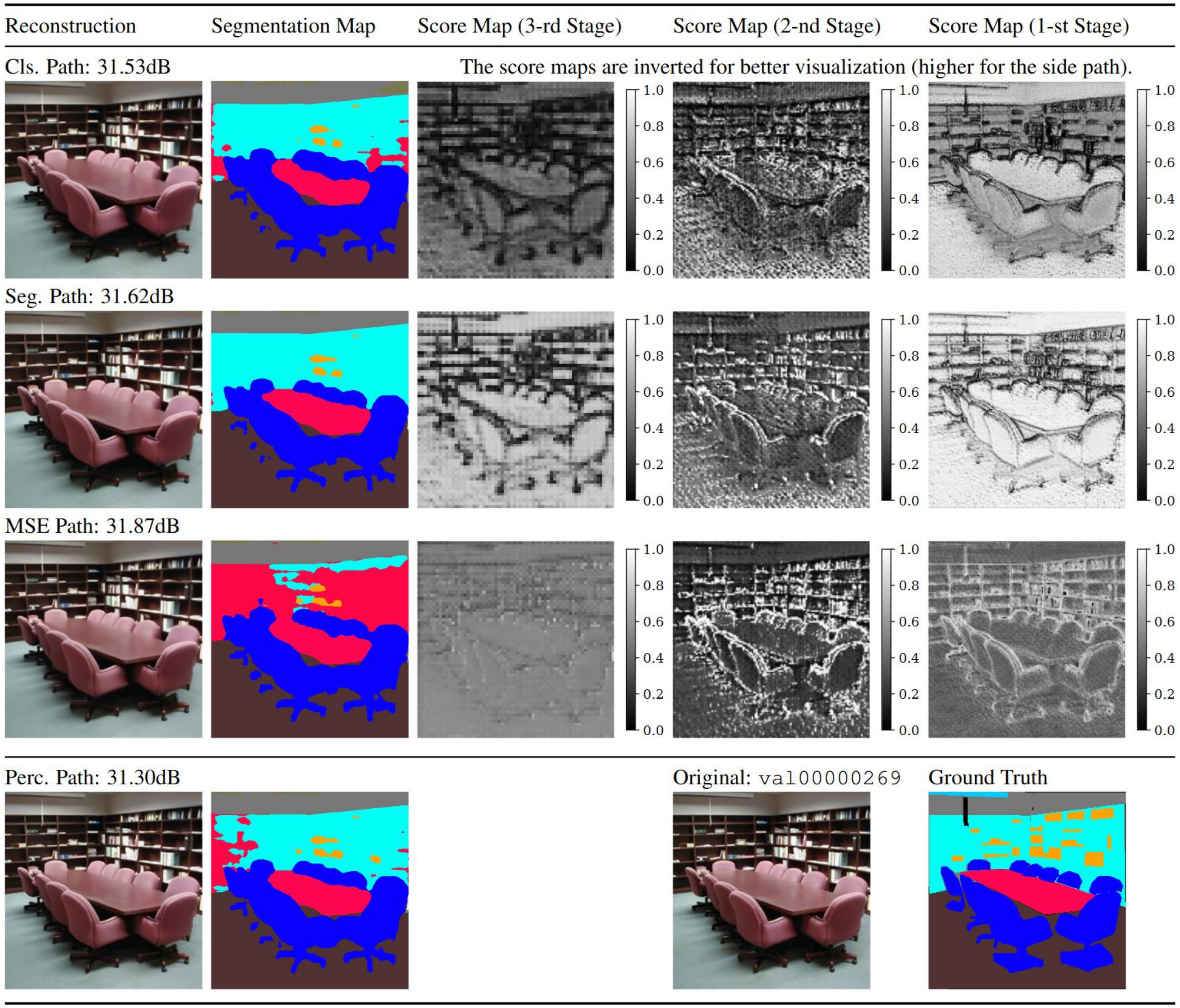
Visualization

Original	MRIC ($\beta = 2.56$)	MPA ($\alpha = 0.0$)	MPA ($\alpha = 0.0001$)	MPA ($\alpha = 0.001$)	MPA ($\alpha = 0.01$)	MPA ($\alpha = 0.1$)	MPA ($\alpha = 1.0$)	MRIC ($\beta = 0.0$)
3f273	0.0540bpp 31.16dB	0.0509bpp 30.00dB	0.0509bpp 30.07dB	0.0509bpp 30.50dB	0.0509bpp 31.08dB	0.0509bpp 31.32dB	0.0509bpp 31.40dB	0.0540bpp 32.17dB
								
88c58	0.0475bpp 32.29dB	0.0459bpp 31.35dB	0.0459bpp 31.42dB	0.0459bpp 31.82dB	0.0459bpp 32.52dB	0.0459bpp 32.82dB	0.0459bpp 32.94dB	0.0475bpp 33.73dB
								
1487a	0.0755bpp 29.36dB	0.0636bpp 28.44dB	0.0636bpp 28.48dB	0.0636bpp 28.76dB	0.0636bpp 29.52dB	0.0636bpp 30.01dB	0.0636bpp 30.09dB	0.0755bpp 30.96dB
								
f5003	0.0750bpp 30.34dB	0.0660bpp 29.11dB	0.0660bpp 29.14dB	0.0660bpp 29.40dB	0.0660bpp 30.15dB	0.0660bpp 30.48dB	0.0660bpp 30.61dB	0.0750bpp 31.71dB
								

Visualization



Visualization



Diving into MPA

Table 1: Effects of path complexity

MLP Type	ϕ_{MSE} (BD-Rate ↓)	ϕ_{cls} (Acc. ↑)	ϕ_{seg} (mIoU ↑)
Bottleneck	19.51%	76.72%	37.41%
Inv. Bottleneck	16.04%	77.16%	37.76%

Table 2: Cross-validations on path choices

Task (Metric)	ϕ_{perc}	ϕ_{MSE}	ϕ_{cls}	ϕ_{seg}
MSE (BD-Rate ↓)	49.61%	**16.04%**	32.81%	34.19%
Cls. (Acc. ↑)	76.66%	60.59%	**76.77%**	73.57%
Seg. (mIoU ↑)	36.17%	28.34%	35.34%	**37.41%**

Table 3: Ablations on encoder

Components	BD-Rate ↓ against VTM
Full MPA	16.04%
w/o Predictors	16.25%
w/o ϕ_{hq}	17.05%
w/o ϕ_{lq}	17.18%

Table 4: Comparison of complexity

Models	#Param.	KFLOPs per pixel	Latency (ms)
MRIC [1]	69.14M	1118.17	11.89
TinyLIC [2]	28.46M	439.29	12.68
+ MLPs	+0.51M∼+2.04M	-56.68∼+0	-0.33∼+0
+ Predictors	+0.03M	+2.23	+0.09
+ Ⓢ & Ⓐ	+0	+0	+0.28
MPA	29.00M∼30.53M	384.84∼441.52	12.72∼13.05

[1] E. Agustsson, et al. Multi-Realism Image Compression With a Conditional Generator. In CVPR, 2023.
[2] M. Lu, et al. High-Efficiency Lossy Image Coding through Adaptive Neighborhood Information Aggregation. arXiv preprint arXiv:2204.11448, 2022.

Conclusion

- Key Contribution:
  - Introduces the Multi-Path Aggregation (MPA) method for unified image coding that supports both human and machine vision tasks.
- Advantages:
  - Achieves comparable performance to SOTA single-task models with reduced parameter and computational costs.
  - Supports flexible task expansion with minimal parameter fine-tuning.
- Performance Highlights:
  - Demonstrates strong rate-distortion and rate-perception performance.
  - Shows excellent results in machine vision tasks like classification and segmentation, achieving near full-tuning models.
- Applications and Future Work:
  - Suitable for multi-task scenarios that require simultaneous human viewing and machine analysis.
  - Future work can focus on joint multi-path optimization to further improve performance and reduce latency.

Thank you!

Xu Zhang, Peiyao Guo, Ming Lu, Zhan Ma

Vision Lab, Nanjing University