Synonymous Variational Inference for Perceptual Image Compression

Zijian Liang, Kai Niu, Changshuo Wang, Jin Xu, Ping Zhang. (Accepted by ICML 2025 Poster)

Presenter: Zijian Liang
`liang1060279345@bupt.edu.cn`

PhD Student
Key Laboratory of Universal Wireless Communications, Ministry of Education
Beijing University of Posts and Telecommunications

June 15, 2025

Bacckground: The Rate-Distortion-Perception tradeoff

Blau and Michaeli demonstrated the apparent tradeoff between the perceptual quality and the distortion measure that widely exists in various distortion measures [1] (CVPR2018), and extended the classic rate-distortion tradeoff to a triple tradeoff version [2] (ICML2019):

$$R(D,P) = \min_{p(\hat{\boldsymbol{x}}|\boldsymbol{x})} I\left(\boldsymbol{X};\hat{\boldsymbol{X}}\right) \quad \text{s.t.} \quad \mathbb{E}_{\boldsymbol{x},\hat{\boldsymbol{x}}\sim p(\boldsymbol{x},\hat{\boldsymbol{x}})}\left[d\left(\boldsymbol{x},\hat{\boldsymbol{x}}\right)\right] \leq D, \quad d_p\left(p_{\boldsymbol{x}}, p_{\hat{\boldsymbol{x}}}\right) \leq P. \tag{1}$$

- They define the perceptual quality index $d_p\left(p_{\boldsymbol{x}}, p_{\hat{\boldsymbol{x}}}\right)$ based on some divergence between distributions of the source and reconstructed images (supported by GAN-based schemes).
- It surpasses the support of Shannon's classical information theory, thus shifting our focus to **semantic information theory** [3] that focuses on higher-level information processing [4].

Figure: An example presented by HifiC [5] (NeurIPS2020)

A semantic information viewpoint

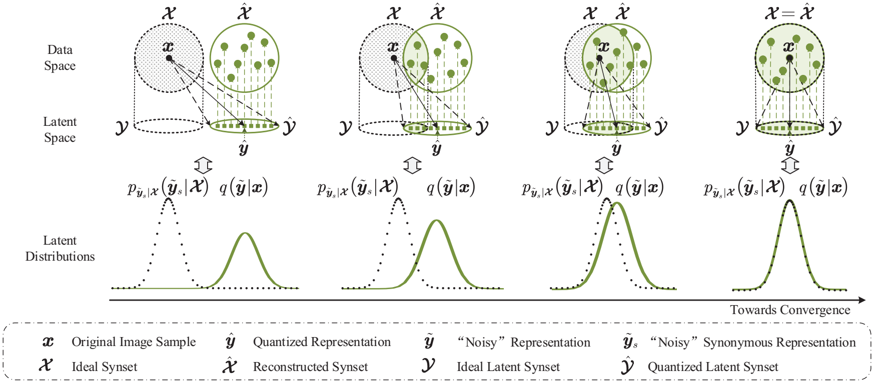

Figure: An illustration of the optimization direction for Synonymous Variational Inference.

- **Basic assumption**: **Images in an ideal synset $\mathcal{X}$ sharing the same latent synonymous representation $\boldsymbol{y}_s$.**
- **Optimization direction**: Minimizing a partial semantic KL divergence, i.e., $\min \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})} D_{\mathsf{KL},s}\left[q||p_{\tilde{\boldsymbol{y}}_s|\boldsymbol{\mathcal{X}}}\right]$.

Synonymous Variational Inference: Derivation

北京郵電大學 Beijing University of Posts and Telecommunications

Lemma

When the source considers the existence of an ideal synset $\boldsymbol{\mathcal{X}}$ and the decoder places the reconstructed sample in a reconstructed synset $\tilde{\boldsymbol{\mathcal{X}}}$, the minimization of the expected negative log synonymous likelihood term

$$\min \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\mathbb{E}_{\tilde{\boldsymbol{y}}\sim q}\left[-\log p_{\boldsymbol{\mathcal{X}}|\tilde{\boldsymbol{y}}_s}\left(\boldsymbol{\mathcal{X}}|\tilde{\boldsymbol{y}}_s\right)\right]$$
$$\Longleftrightarrow \min \lambda_d \cdot \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\mathbb{E}_{\tilde{\boldsymbol{y}}\sim q}\mathbb{E}_{\tilde{\boldsymbol{x}}_i\in\tilde{\boldsymbol{\mathcal{X}}}|\tilde{\boldsymbol{y}}_s}\left[d\left(\boldsymbol{x},\tilde{\boldsymbol{x}}_i\right)\right] + \lambda_p \cdot \mathbb{E}_{\tilde{\boldsymbol{y}}\sim q}\mathbb{E}_{\tilde{\boldsymbol{x}}_i\in\tilde{\boldsymbol{\mathcal{X}}}|\tilde{\boldsymbol{y}}_s} D_{\mathsf{KL}}\left[p_{\boldsymbol{x}}||p_{\tilde{\boldsymbol{x}}_i}\right], \quad (2)$$

in which λ_d and λ_p are the tradeoff factors for the expected distortion (typically expected means-squared error, i.e., E-MSE) term and the expected KL divergence (E-KLD) term, respectively.

By using the proposed SVI, i.e., minimizing the partial semantic KL divergence given in (**??**),

$$\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})} D_{\mathsf{KL},s}\left[q||p_{\tilde{\boldsymbol{y}}_s|\boldsymbol{\mathcal{X}}}\right] = \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\mathbb{E}_{\tilde{\boldsymbol{y}}\sim q}\Big[\underbrace{\log q(\tilde{\boldsymbol{y}}|\boldsymbol{x})}_{\to 0}\underbrace{-\log p_{\boldsymbol{\mathcal{X}}|\tilde{\boldsymbol{y}}_s}\left(\boldsymbol{\mathcal{X}}|\tilde{\boldsymbol{y}}_s\right)}_{\text{Tradeoff in Lemma}}\underbrace{-\log p_{\tilde{\boldsymbol{y}}_s}\left(\tilde{\boldsymbol{y}}_s\right)}_{\text{Rate}}\Big] + \text{const.} \quad (3)$$

This target corresponds to a ***Synonymous Rate-Distortion-Perception Tradeoff***, which can be shown as

$$\mathcal{L}_{\boldsymbol{\mathcal{X}}} = \underbrace{\lambda_r \cdot \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[-\log p_{\hat{\boldsymbol{y}}_s}\left(\hat{\boldsymbol{y}}_s\right)\right]}_{\text{Synonymous Coding Rate}} + \underbrace{\lambda_d \cdot \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\mathbb{E}_{\hat{\boldsymbol{x}}_i\in\hat{\boldsymbol{\mathcal{X}}}|\hat{\boldsymbol{y}}_s}\left[d\left(\boldsymbol{x},\hat{\boldsymbol{x}}_i\right)\right]}_{\text{Expected Distortion}} + \underbrace{\lambda_p \cdot \mathbb{E}_{\hat{\boldsymbol{x}}_i\in\hat{\boldsymbol{\mathcal{X}}}|\hat{\boldsymbol{y}}_s} D_{\mathsf{KL}}\left[p_{\boldsymbol{x}}||p_{\hat{\boldsymbol{x}}_i}\right]}_{\text{Expected KL Divergence (Perception)}}, \quad (4)$$

Compatibility Analysis

- ***Compatibility with Existing Rate-Distortion-Perception Tradeoff:*** When the reconstructed synset is not considered (equal to the reconstructed synset contains only one sample, represented as $\hat{\boldsymbol{\mathcal{X}}} = \{\hat{\boldsymbol{x}}\}$), the optimization objective will be degraded into the existing rate-distortion-perception tradeoff:

$$
\begin{aligned}
R(\boldsymbol{\mathcal{X}}) = \min_{p(\hat{\boldsymbol{\mathcal{X}}}|\boldsymbol{x})} & \; I\left(\boldsymbol{X}; \mathring{\hat{\boldsymbol{X}}}\right) \\
\text{s.t.} \quad & \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})} \mathbb{E}_{\hat{\boldsymbol{x}}_i \in \hat{\boldsymbol{\mathcal{X}}}|\hat{\boldsymbol{y}}_s} \left[d\left(\boldsymbol{x}, \hat{\boldsymbol{x}}_i\right)\right] \leq D, \\
& \mathbb{E}_{\hat{\boldsymbol{x}}_i \in \hat{\boldsymbol{\mathcal{X}}}|\hat{\boldsymbol{y}}_s} D_{\mathsf{KL}} \left[p_{\boldsymbol{x}} || p_{\hat{\boldsymbol{x}}_i} \right] \leq P,
\end{aligned}
\quad \xrightarrow{\hat{\boldsymbol{\mathcal{X}}} = \{\hat{\boldsymbol{x}}\}} \quad
\begin{aligned}
R(D,P) = \min_{p(\hat{\boldsymbol{x}}|\boldsymbol{x})} & \; I\left(\boldsymbol{X}; \hat{\boldsymbol{X}}\right) \\
\text{s.t.} \quad & \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})} \left[d\left(\boldsymbol{x}, \hat{\boldsymbol{x}}\right)\right] \leq D, \\
& D_{\mathrm{KL}} \left[p_{\boldsymbol{x}} || p_{\hat{\boldsymbol{x}}} \right] \leq P,
\end{aligned}
\tag{5}
$$

- ***Compatibility with Traditional Rate-Distortion Tradeoff:*** When the ideal synset is not considered (equal to the ideal synset contains only the original image, represented as $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}\}$), the expected synonymous likelihood term will be degraded into the usual likelihood term, i.e.,

$$
\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})} \mathbb{E}_{\tilde{\boldsymbol{y}}\sim q} \left[-\log p_{\boldsymbol{\mathcal{X}}|\tilde{\boldsymbol{y}}_{\boldsymbol{s}}} \left(\boldsymbol{\mathcal{X}} | \tilde{\boldsymbol{y}}_{\boldsymbol{s}}\right)\right] \quad \xrightarrow{\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}\}} \quad \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})} \left[-\log p_{\boldsymbol{x}|\tilde{\boldsymbol{y}}} \left(\boldsymbol{x} | \tilde{\boldsymbol{y}}\right)\right]. \tag{6}
$$

Therefore, the relationship with the traditional rate-distortion tradeoff can be represented by

$$
\begin{aligned}
R(\boldsymbol{\mathcal{X}}) = \min_{p(\hat{\boldsymbol{\mathcal{X}}}|\boldsymbol{x})} & \; I\left(\boldsymbol{X}; \mathring{\hat{\boldsymbol{X}}}\right) \\
\text{s.t.} \quad & \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})} \mathbb{E}_{\hat{\boldsymbol{x}}_i \in \hat{\boldsymbol{\mathcal{X}}}|\hat{\boldsymbol{y}}_s} \left[d\left(\boldsymbol{x}, \hat{\boldsymbol{x}}_i\right)\right] \leq D, \\
& \mathbb{E}_{\hat{\boldsymbol{x}}_i \in \hat{\boldsymbol{\mathcal{X}}}|\hat{\boldsymbol{y}}_s} D_{\mathsf{KL}} \left[p_{\boldsymbol{x}} || p_{\hat{\boldsymbol{x}}_i} \right] \leq P,
\end{aligned}
\quad \xrightarrow[(\hat{\boldsymbol{\mathcal{X}}} = \{\hat{\boldsymbol{x}}\})]{\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}\}} \quad
\begin{aligned}
R(D) = \min_{p(\hat{\boldsymbol{x}}|\boldsymbol{x})} & \; I\left(\boldsymbol{X}; \hat{\boldsymbol{X}}\right) \\
\text{s.t.} \quad & \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})} \left[d\left(\boldsymbol{x}, \hat{\boldsymbol{x}}\right)\right] \leq D.
\end{aligned}
\tag{7}
$$

Framework of Synonymous Image Compression (SIC)

北京郵電大學 Beijing University of Posts and Telecommunications

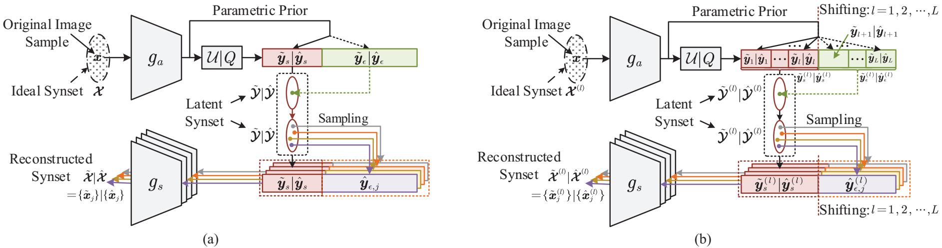

Figure: Processing frameworks of SIC. (a) The general framework. (b) The progressive framework.

We implement a progressive SIC model, and optimize it with a group of loss functions that alternatively train for the level $l = 1, 2, \cdots, L$ step by step, i.e.:

$$\mathcal{L}^{(l)} = \alpha \mathcal{L}_{\boldsymbol{\mathcal{X}}}^{(l)} + (1-\alpha)\,\mathcal{L}_{\boldsymbol{\mathcal{X}}}^{(L)} + \mathcal{L}_c^{(l)}, l = 1, 2, \cdots, L, \quad (8)$$

in which $\mathcal{L}_{\boldsymbol{\mathcal{X}}}^{(l)}$ is represented by

$$\mathcal{L}_{\boldsymbol{\mathcal{X}}}^{(l)} = \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[-\lambda_r^{(l)} \cdot \log p_{\hat{\boldsymbol{y}}_s^{(l)}}\left(\hat{\boldsymbol{y}}_s^{(l)}\right) + \frac{1}{M}\sum_{i=1}^{M}\left(\lambda_d^{(l)} \cdot \mathsf{MSE}(\boldsymbol{x}, \hat{\boldsymbol{x}}_i^{(l)}) + \lambda_p^{(l)} \cdot \mathsf{LPIPS}(\boldsymbol{x}, \hat{\boldsymbol{x}}_i^{(l)})\right)\right], \quad (9)$$

Experimental Illustration: Results and Analysis

北京郵電大學 Beijing University of Posts and Telecommunications

We focus on the **DISTS** measure [6], due to its resampling tolerance, which aligns more closely with the human understanding of perceptual similarity, i.e., typified synonymous relationships.

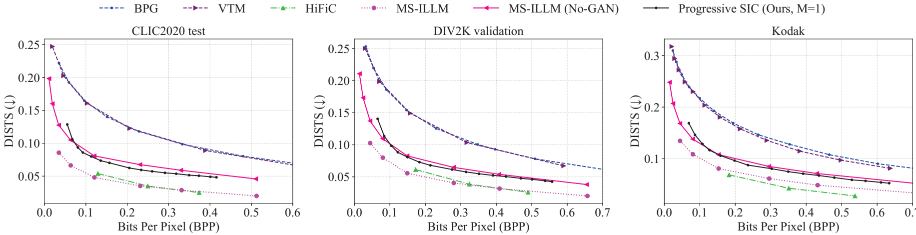

Figure: Comparisons with methods [7, 8, 5, 9] using DISTS on different datasets.

Experimental results show perceptual quality adaptability across various rates **using a single model**, with the perceptual quality of the reconstructed image improving as the coding rate increases.

For the concerned DISTS measure, our method **surpasses the No-GAN MS-ILLM solution (also trained with LPIPS) in a large coding rate range**. This performance is demonstrated under conditions where the PSNR quality continuously approaches and even exceeds the comparison No-GAN schemes, and the LPIPS quality remains very similar, thus verifying a comparable rate-distortion-perception performance.

Experimental Illustration: Results and Analysis

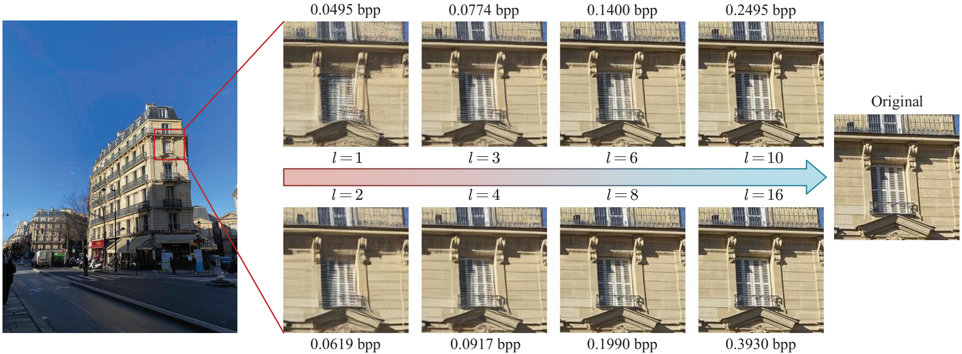

Figure: Visualization results of reconstructed images at different synonymous levels using progressive SIC ($M = 1$). Image from the CLIC2020 test set [10].

- Low synonymous levels → Low coding rates → Large Synset → Focus more on global content semantic;
- High synonymous levels → High coding rates → Small Synset → Focus more on local detail semantic.

Main References

北京邮电大学 Beijing University of Posts and Telecommunications

Y. Blau and T. Michaeli, “The perception-distortion tradeoff,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

——, “Rethinking lossy compression: The rate-distortion-perception tradeoff,” in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, K. Chaudhuri and R. Salakhutdinov, Eds. PMLR, 2019, pp. 675–685.

K. Niu and P. Zhang, “A mathematical theory of semantic communication,” *Journal on Communications*, vol. 45, no. 6, pp. 7–59, 2024. [Online]. Available: https://www.joconline.com.cn/en/article/doi/10.11959/j.issn.1000-436x.2024111/

W. Weaver, “Recent contributions to the mathematical theory of communication,” *ETC: a review of general semantics*, pp. 261–281, 1953.

F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, “High-fidelity generative image compression,” in *Advances in Neural Information Processing Systems (NeurIPS)*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds. Curran Associates, Inc., 2020, pp. 11 913–11 924.

K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, “Image quality assessment: Unifying structure and texture similarity,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 5, pp. 2567–2581, 2020.

F. Bellard, “BPG image format,” 2015. [Online]. Available: https://bellard.org/bpg

“VVCSoftware_VTM,” Version 23.4. [Online]. Available: https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM.git

M. J. Muckley, A. El-Nouby, K. Ullrich, H. Jegou, and J. Verbeek, “Improving statistical fidelity for neural image compression with implicit local likelihood models,” in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds. PMLR, 2023, pp. 25 426–25 443.

G. Toderici, L. Theis, N. Johnston, E. Agustsson, F. Mentzer, J. Ballé, W. Shi, and R. Timofte, “CLIC 2020: Challenge on learned image compression, 2020,” 2020.

Thank you for your attention!