

# JPD-SE: High-Level Semantics for Joint Perception-Distortion Enhancement in Image Compression

Shiyu Duan\*, Huaijin Chen, and Jinwei Gu

**Abstract**—While humans can effortlessly transform complex visual scenes into simple words and the other way around by leveraging their high-level understanding of the content, conventional or the more recent learned image compression codecs do not seem to utilize the semantic meanings of visual content to their full potential. Moreover, they focus mostly on rate-distortion and tend to underperform in perception quality especially in low bitrate regime, and often disregard the performance of downstream computer vision algorithms, which is a fast-growing consumer group of compressed images in addition to human viewers. In this paper, we (1) present a generic framework that can enable any image codec to leverage high-level semantics and (2) study the joint optimization of perception quality and distortion. Our idea is that given *any* codec, we utilize high-level semantics to augment the low-level visual features extracted by it and produce essentially a new, semantic-aware codec. We propose a three-phase training scheme that teaches semantic-aware codecs to leverage the power of semantic to jointly optimize rate-perception-distortion (R-PD) performance. As an additional benefit, semantic-aware codecs also boost the performance of downstream computer vision algorithms. To validate our claim, we perform extensive empirical evaluations and provide both quantitative and qualitative results.

**Index Terms**—image compression, high-level semantics, generative adversarial networks

## I. INTRODUCTION

THE pervasive use of mobile devices equipped with powerful cameras has pushed the generation of high-definition images and videos to an unprecedented rate. Such constant and ubiquitous data collection poses an urgent motivation for better visual data compression techniques.

Consider how humans compress visual input (Fig. 1): The sender first comprehends the visual content and then extracts high-level semantic descriptions. He or she then describes them in a few words and the receiver, upon obtaining this information, can reconstruct visually complex scenes. The fact that we can all leverage our understanding and associate concise high-level concepts such as “boat” with pixel intensities and arrangements enables us to condense even tens of millions of pixels into several words and the other way around. Being able to effortlessly leverage high-level meanings of visual content to increase the efficiency of its processing has been a long-sought goal since at least the advent of MPEG-7 [1].

SD (michaelshiyu3@gmail.com) is with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, 32611. HC (chenhuaijin@sensebrain.site) and JG (gujinwei@sensebrain.site) are with SenseBrain Technology, San Jose, CA, 95131.

\*Work done during internship at SenseBrain.

Uncompressed Inputs	Compressed Representations	Decompressed Outputs	Metrics
	Semantics Only “A sailboat in the water.” (Humans)		Perception Quality: Good 😊 Accuracy: Low 😞 Distortion: High 😞
	Semantics Only “A small boat in a body of water.” (NIC → AttnGAN)		Perception Quality: Medium 😐 Accuracy: Low 😞 Distortion: High 😞
	Low-Level Visuals Only  (Current Codecs)		Perception Quality: Medium 😐 Accuracy: Medium 😐 Distortion: Good 😊
	Semantics + Low-Level Visuals  (Our Method)		Perception Quality: Good 😊 Accuracy: Good 😊 Distortion: Good 😊

Fig. 1: Human can leverage high-level semantics to compress efficiently (top row). While this behavior can be replicated by deep learning models [20], [21] (the second row from the top), existing compression codecs underutilize semantics (the third row). We enhance existing codecs by augmenting the low-level visuals with semantics and consequently let the codecs leverage a high-level “understanding” to compress better along three performance axes: perception quality [22], performance of downstream vision algorithms (accuracy), and distortion (bottom row).

However, efforts in this direction have seen limited success due to the sheer difficulties in computationally implementing how humans extract and leverage the high-level meanings of low-level visual features. As of today, existing image compression algorithms still underutilize high-level semantics [2]–[19].

Another limitation in existing codec design paradigms is that the models are typically optimized only for rate-distortion (R-D), yet the traditional R-D metrics are insufficient especially in low bitrate [22], [23], causing the codecs to produce visually unappealing results in low-rate compression.<sup>1</sup> Apart from human viewers, existing codecs also ignore “machine

<sup>1</sup>As a sign of the image compression community realizing the severity of this issue, the low-rate track of the 2020 Workshop and Challenge on Learned Image Compression (CLIC) is for the first time using human ratings over PSNR/MS-SSIM [24] as the primary metric [25].

viewers”, i.e., the downstream computer vision algorithms performing tasks such as face recognition, object tracking, etc. on the decoded content. With the fast-expanding use of machine learning tools in image understanding tasks, rate-accuracy (R-A), where “accuracy” refers to the performance of downstream machine viewers, must become another important axis when evaluating image codec.

We propose a universal framework to achieve joint perception-distortion optimization with *any* given image codec through semantic enhancement (JPD-SE). Given any image codec, the idea is to augment its hidden representations constructed from low-level visuals with high-level semantics. Then a synthesis network leverages this semantics to enhance the codec output (Fig. 3). A three-stage training scheme is used to teach the model to best leverage semantics to achieve joint perception-distortion optimization.

By simply feeding semantics to the codec, we give the codec more leverage to improve perception-distortion-accuracy performance. First, just like humans know that the sky is usually blue and grass green, a semantic-aware codec can learn what low-level features are more likely to appear in different semantic regions. This “prior knowledge” enables the decoder to create more natural and perceptually pleasing reconstructions. This also helps remove noise on the low-level visual features without training on noisy examples since noise can be considered as unnatural visuals. Second, since complex low-level features can often be packed into concise high-level semantic concepts, more information can be conveyed using a blend of semantics and low-level visuals than using only the latter. Therefore, a semantic-aware codec can leverage this extra degree of freedom and save bits. Finally, this high-level comprehension of the image helps the codec retain important semantic information that is critical for downstream computer vision tasks such as object detection or tracking.

To teach the model to utilize semantics, we propose a three-stage training scheme aiming at jointly optimize perception-distortion performance. The first stage trains the model with a GAN loss, guiding it to hallucinate photorealistic pixels from semantics. The second stage combines the GAN loss with a distortion loss to jointly improve perceptual quality and distortion. Finally, training with both GAN and distortion empirically leads to models that overly emphasize on perceptual quality, generating hallucination artifacts. Therefore, we train for a third stage with distortion loss alone, producing better joint perception-distortion performance. Although our training does not directly optimize rate-accuracy performance, we demonstrate through a posteriori evaluation that semantically-enhanced codecs can boost the performance of downstream vision algorithms.

We test on Cityscapes [26] and ADE20k [27], [28] at 1024×512 resolution with learned and conventional codecs (JPEG, JPEG 2000 [4], WebP [2], and BPG [3]) as the backbone image codec and semantic and instance segmentation maps as the semantics. We both quantitatively and qualitatively show that our semantically-enhanced (SE) codecs achieve favorable rate-perception-accuracy-distortion performance as quantified via a large-scale user study, accuracy of a downstream bounding-box object detector, and three distortion

measures (LPIPS [29], MS-SSIM [24], and PSNR). Extensive ablation study was performed to validate the usefulness of the proposed components. In particular, we demonstrate that the improvements are beyond what could be achieved with a post-processing network using purely low-level visuals, proving the value of semantics.

To sum up, the main contributions of this work include:

- We conduct a systematic study on the efficacy of semantics as a fundamental supplement to visuals in the general image compression setting.
- We propose a generic framework that enables any given codec, learned or hand-crafted, to leverage high-level semantics. As part of this framework, a three-stage training scheme guides the codec to leverage semantics to jointly optimize perception-distortion.
- We evaluate perception-accuracy-distortion performance on full-resolution images quantified by a large-scale user study, accuracy of downstream bounding-box object detection, and three distortion metrics.
- We hypothesize and verify that enhancing codecs with high-level semantics improves perception-accuracy-distortion.

Code is available at <https://github.com/SenseBrain/JPD-SE>.

## II. RELATED WORK

*a) Image Compression:* The key components of an image codec include an encoder, which transforms the original image into a more compressible representation, and a decoder, which reconstructs the image from a possibly quantized version of this new representation. Traditionally, these components were hand-crafted by experts. Some commonly used engineered image codecs include JPEG, JPEG 2000 [4], WebP [2], BPG [3], PNG [30], and FLIF [31]. PNG and FLIF are designed only for lossless compression. Many works on learned compression have emerged over recent years and the proposed codecs with learnable components produced favorable results compared to the traditional ones but are conceptually much simpler due to their end-to-end nature [5], [7]–[18]. Our proposed method can work with any learned or engineered image codec, leveraging high-level semantics to improve compression quality as measured by joint PAD performance.

*b) Semantics in Image Compression:* Most existing codecs do not explicitly utilize high-level semantics during encoding or decoding. [14], [32] proposed methods for content-weighted bitrate control without explicitly utilizing high-level semantics.

[6], [33] explored explicitly utilizing semantics in image compression but both to a limited extent. Specifically, [6] only used semantics for bitrate allocation in a somewhat constrained setting, requiring that users choose to preserve some semantic regions while ignoring others. [33] used semantics only for up/downsampling. Also, BPG was used to compress the residuals of its semantics-aware codec, making it difficult to assess the contributions from semantics. Apart from the different compression objectives (to be discussed in the next section), our work differs from the previous two in the following aspects. First, we study how semantics can serve



Fig. 2: Semantically-enhanced (SE) codecs vs. originals. Note how the SE codecs are able to “understand” an input image and produce a more natural and visually appealing reconstruction in places where the originals failed (e.g., the boundary between the tires and the road). Fig. 2k was produced by the same SE codec that produced 2j using dummy semantics during test (-s; test), illustrating the importance of semantics. Shown are cropped images. Full images are in the supplementary materials.

as a fundamental supplement to visuals that boosts PAD performance instead of only as some auxiliary side information. Second, our work considers the general compression setting without any limiting assumption such as that distinct semantic regions are of different levels of importance to the user, which is essential to [6].

*c) Compression Objective:* Most existing works focus only on R-D and ignore perception quality [2]–[5], [7]–[18], [33]. Codecs in this category fail to produce reconstructions that look good to humans in low bitrates despite their strong R-D performance, as observed in previous work [6]. The underlying reason of this mismatched performance is that distortion, including the more recent “perceptual” distortions [24], [29], [34], is fundamentally at odds with perception quality especially in low bitrate settings [23]. [6] focuses on R-P while ignoring distortion completely. The reconstructions look appealing at the first glance but have low fidelity. This suggests that when designing codecs, we should simultaneously consider both perception and distortion for better overall quality. [17], [35] focus on joint perception-distortion optimization, but only on thumbnail images.

To the best of our knowledge, we are the first to focus on jointly optimizing perception quality and distortion on full-resolution images. We quantify perception quality directly through a user study that has the largest scale among existing studies in terms of the number of codecs tested, the range of bitrates covered, and the number of users involved.

We also demonstrate a new approach for optimizing perception quality. Existing works typically add a quantification term for perception in the form of some GAN loss to the overall objective function [6], [17], [35]. We show that sending

semantics to the decoder naturally improves perception quality.

In addition to perception and distortion, we propose to evaluate codec performance along a third “accuracy” axis. This new performance axis is becoming increasingly important due to the wide use of vision algorithms on compressed content. Note that apart from a clear tradeoff between rate and “accuracy” (Fig. 6), we do not claim that a tradeoff necessarily exists among “accuracy” and perception or distortion.

*d) Semantic Image Synthesis:* Semantic image synthesis methods aim at synthesizing photorealistic and semantically consistent images from given semantic descriptions such as text, sketches, and semantic segmentation maps [21], [36]–[46]. This problem can be modeled and subsequently solved with a conditional GAN as follows. Given semantics  $\mathbf{x}_{\text{sem}}$ , one would like a generator  $G$  that can generate an image  $G(\mathbf{x}_{\text{sem}})$  that has the given semantics and is distributed according to the distribution of some real image  $\mathbf{x}$ . This goal can be achieved by training  $G$  (and a discriminator  $D$ ) with the following conditional GAN objective function:

$$L_{\text{GAN}} = \mathbb{E} f(D(\mathbf{x}, \mathbf{x}_{\text{sem}})) + \mathbb{E} g(D(G(\mathbf{x}_{\text{sem}}), \mathbf{x}_{\text{sem}})), \quad (1)$$

where  $f$  and  $g$  depend on the GAN formulation used.

Recent methods have been able to synthesize high-definition photographic images with pixel-level semantic accuracy from semantic segmentation maps using conditional GANs [42]–[46]. Instance-wise low-level visual features are sometimes used to enhance the realism of the synthesized images [43]. For semantic image synthesis, there is no “original image” and the evaluation standards are completely different: The generated images are typically evaluated in terms of perception quality and semantic consistency. In this work, we consider

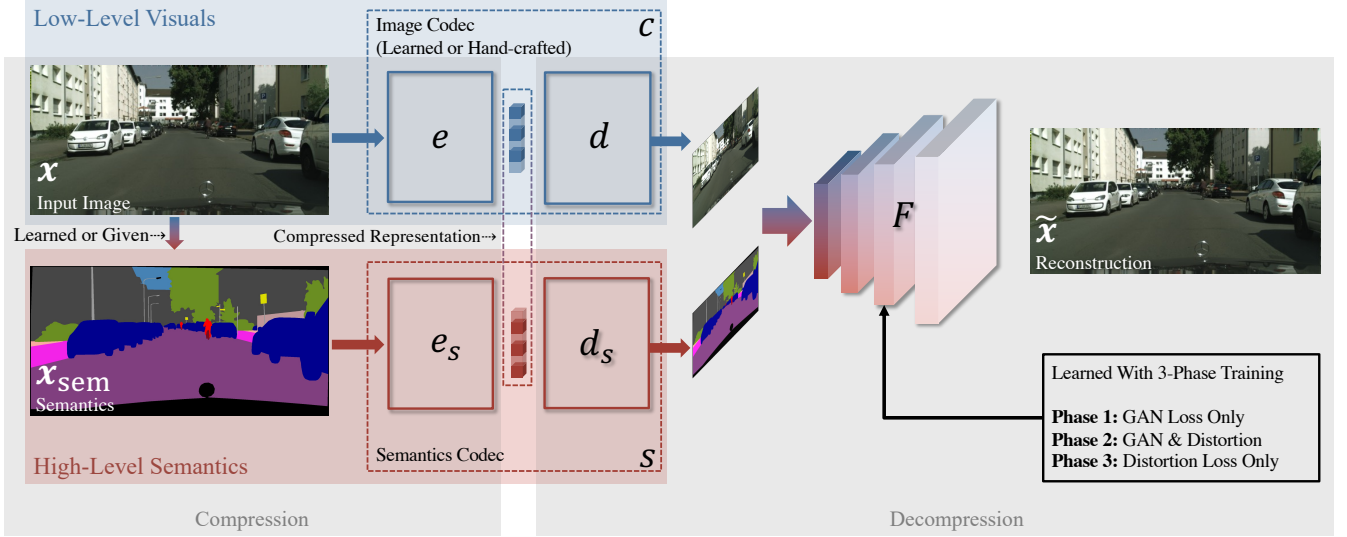


Fig. 3: Our proposed framework consists of a given backbone image codec  $c$ , a semantics codec  $s$ , and a fusion network  $F$ . To compress, we use  $e$  and  $e_s$  together as the encoder that constructs a hidden representation of the image with both high-level semantics and low-level visuals. To decompress,  $d$ ,  $d_s$ , and  $F$  work jointly as the decoder.  $d$  and  $d_s$  reconstruct visuals and semantics from the compressed representation, respectively.  $F$  fuses their outputs to produce the final result.  $F$  is learned, whereas the other modules are not necessarily so. In particular, we tested with both learned and engineered codecs as the backbone  $c$ . For the learned components, a three-phase training scheme is proposed to effectively optimize for both distortion and perception quality. We focus on segmentation maps as semantics for this work although the framework is generic.

rate, perception quality, performance of downstream vision algorithms, and distortion with respect to the original image.

### III. BUILDING SEMANTICALLY-ENHANCED CODECS

1) *The Main Pipeline:* Denote the input image as  $\mathbf{x} \in X$  and its semantics  $\mathbf{x}_{\text{sem}} \in X_{\text{sem}}$ , where  $X$  is a space of image tensors with shape  $H \times W \times C$  and  $X_{\text{sem}}$  a space of semantics. The semantics can be obtained computationally with another component that may be jointly optimized with our compression pipeline.

**Backbone codec  $c$ :** As the core of the pipeline, we need a backbone image codec  $c = d \circ e$  that encodes and decodes the image input, where  $e : X \rightarrow Z$  is an encoder,  $d : Z \rightarrow X$  a decoder,  $\circ$  denotes function composition, and  $e(\mathbf{x}) \in Z$  is a hidden representation of  $\mathbf{x}$  in some feature space  $Z$  that will be transmitted after potential quantization and entropy coding. For example, this backbone codec can be JPEG or a learned image codec. And we evaluate different choices of  $c$  in experiments.

**Semantics codec  $s$ :** Concurrent to the backbone image codec  $c$ , we also need a semantics codec  $s = d_s \circ e_s$  that is responsible for encoding and decoding semantics.  $e_s : X_{\text{sem}} \rightarrow Z_{\text{sem}}$  here is an encoder,  $d_s : Z_{\text{sem}} \rightarrow X_{\text{sem}}$  a decoder, and  $e_s(\mathbf{x}_{\text{sem}})$  a hidden representation of  $\mathbf{x}_{\text{sem}}$ .  $Z_{\text{sem}}$  is a feature space for semantics. We shall present a non-learned component that serves as  $s$  for encoding and decoding segmentation maps in Section III-3.

**Fusion network  $F$ :** On top of image codec  $c$  and semantics codec  $s$ , we need another fusion network  $F : X \times X_{\text{sem}} \rightarrow X$  that synthesizes the reconstructed image  $\tilde{\mathbf{x}}$  by fusing together the outputs of the two codecs.  $F$  should leverage semantics to

enhance the output from the backbone codec. Mathematically,  $F$  maps the tuple  $(c(\mathbf{x}), s(\mathbf{x}_{\text{sem}}))$  to  $\tilde{\mathbf{x}}$ . We implement  $F$  using a conditional GAN generator.

Concisely, the end-to-end pipeline can be described as follows.

$$\left. \begin{array}{l} \mathbf{x} \xrightarrow{c} c(\mathbf{x}) \\ \mathbf{x}_{\text{sem}} \xrightarrow{s} s(\mathbf{x}_{\text{sem}}) \end{array} \right\} (c(\mathbf{x}), s(\mathbf{x}_{\text{sem}})) \xrightarrow{F} \tilde{\mathbf{x}}$$

Fig. 3 gives a schematic illustration of this framework. When  $c$  is a learned codec, it may be included in the joint optimization. The proposed framework enjoys wide applicability since any existing image codec can be plugged into the pipeline as the backbone  $c$  and be enhanced semantically.

Note that high-level semantics can take many forms. For example, one can use a class segmentation map, an instance segmentation map, sketches, a set of bounding boxes, text descriptions and so on. And each form potentially requires an architecturally different  $s$ .

**Loss Function Components:** The main components of the training objective function include a distortion term  $L_{\text{dist}}$ , a GAN term  $L_{\text{GAN}}$ , and optionally, a rate term. The GAN term facilitates the reconstruction of image from a blend of semantics and visuals. And the distortion term ensures that the reconstruction has high fidelity to the original. The GAN term can be interpreted as quantifying perception quality as it can be viewed as a divergence measure between the densities of  $\tilde{\mathbf{x}}$  and  $\mathbf{x}$  [47]. We do not include a loss term that optimizes accuracy directly. Instead, through a posteriori evaluation, we demonstrate that the utilization of semantics improves rate-accuracy performance as an additional benefit.



When using  $L^1$  loss as the distortion term, we have

$$L_{\text{dist}}(\tilde{\mathbf{x}}, \mathbf{x}) = \|\tilde{\mathbf{x}} - \mathbf{x}\|_1. \quad (2)$$

Some other valid choices for  $L_{\text{dist}}$  include the  $L^2$  loss (essentially optimizing PSNR directly), or MS-SSIM [24]. The GAN term is based on the standard conditional GAN formulation Eq. 3:

$$L_{\text{GAN}}(\tilde{\mathbf{x}}, \mathbf{x}) = \mathbb{E} \log(D(\mathbf{x}, \mathbf{x}_{\text{sem}})) + \mathbb{E} \log(1 - D(\tilde{\mathbf{x}}, \mathbf{x}_{\text{sem}})), \quad (3)$$

where  $D$  is the discriminator that maximizes this loss term during training while  $F$  minimizes it. The expectation is estimated through mini batch sample mean.

To stabilize training and enhance visual quality, one typically supplements GAN loss with auxiliary loss terms. VGG [34] loss computes feature space distance and is formulated as

$$L_{\text{VGG}}(\tilde{\mathbf{x}}, \mathbf{x}) = \sum_{i=1}^N \frac{1}{M_i} \left\| V^{(i)}(\mathbf{x}) - V^{(i)}(\tilde{\mathbf{x}}) \right\|_1, \quad (4)$$

where  $V^{(i)}$  denotes the  $i$ th layer with  $M_i$  elements in the VGG network with a total of  $N$  layers [48]. The GAN feature matching loss [43] is formulated similarly and has been shown to lead to improved quality when used alongside the VGG loss.

$$L_{\text{FM}}(\tilde{\mathbf{x}}, \mathbf{x}) = \sum_{i=1}^{N^{(D)}} \frac{1}{M_i^{(D)}} \left\| D^{(i)}(\mathbf{x}) - D^{(i)}(\tilde{\mathbf{x}}) \right\|_1, \quad (5)$$

where  $D^{(i)}$  is the  $i$ th layer with  $M_i^{(D)}$  elements in the GAN discriminator with a total of  $N^{(D)}$  layers. The feature matching loss differs from the VGG loss mainly in that the former computes distance in the feature space of the GAN's discriminator, whereas the latter does so in that of a pre-trained, off-the-shelf VGG network. We used both VGG loss and GAN feature matching loss for our experiments.

To best optimize for joint perception-distortion performance, we empirically found it ideal to train the model in separate stages with each stage involving only some loss terms instead of in one stage with all loss terms. The strategy is described in the following subsection.

2) *A Three-Phase Training Scheme*: To have  $F$  successfully learn how to synthesize visuals from a combination of semantics and visuals, one may begin with training it to minimize the GAN loss  $L_{\text{GAN}}$  supplemented with the VGG loss  $L_{\text{VGG}}$  and the feature matching loss  $L_{\text{FM}}$ , which represents optimizing perceptual quality. Once  $F$  is capable of generating reasonably realistic visual content, it should then be optimized with respect to a weighted sum of the above loss terms and another distortion loss directly on the reconstructed pixels  $L_{\text{dist}}$  such that  $\tilde{\mathbf{x}}$  will be close to  $\mathbf{x}$ . In addition, the combination of the distortion loss  $L_{\text{dist}}$  and the GAN loss  $L_{\text{GAN}}$  stabilizes training since the former can be interpreted as a prior that prevents GAN from mode collapse [6]. Finally, to mitigate the issue that GAN loss usually dominates distortion loss when used together, leading to hallucination artifacts (indicating that the model is overly optimizing perception) [6], we finish training by optimizing  $F$  with respect to  $L_{\text{dist}}$  only. We later show that training for three stages guides the model to produce better

joint perception-distortion quality than training for just the first two. The latter option, as we have observed, led to models that heavily emphasize on perceptual quality but ignore distortion despite careful weighting of the GAN and the distortion loss terms during training. If  $c$  is implemented as a learned codec, it should be jointly optimized with  $F$  during training.

We summarize the three training phases below, in which we ignore the rate term for simplicity.

- Phase 1: Minimize  $L_{\text{GAN}} + \lambda_1 L_{\text{VGG}} + \lambda_2 L_{\text{FM}}$  for some  $\lambda_1, \lambda_2 > 0$ ;
- Phase 2: Minimize  $L_{\text{GAN}} + \lambda_1 L_{\text{VGG}} + \lambda_2 L_{\text{FM}} + \lambda_3 L_{\text{dist}}$  for some  $\lambda_3 > 0$ ;
- Phase 3: Minimize  $L_{\text{dist}}$ .

3) *Compressing the Semantics*: Instead of implementing the semantics encoder  $s$  with yet another network, we propose a simple but effective non-learned approach to better utilize some characteristics of semantics in the form of a segmentation map.<sup>2</sup> We observe that pixel-wise semantics in the form of a segmentation map usually contains large uniform areas separated by a few class or instance boundaries. Instead of compressing the map as a bitmap image, which is prone to creating semantically impossible artifacts such as blurred boundaries, we represent it with a set of paths, each being the boundary of a semantic region. For each path, we additionally store a single number encoding the class or instance identity of the semantic region. And we only keep the two end points for each line segment in the paths. The path-based representation can be obtained from the map with a graph traversal algorithm. Since the boundaries rarely change drastically in natural images, we apply a simple delta encoding on the paths and entropy code the resulting sequences to further reduce the bitrate. This algorithm is lossless but if desired, one can convert it to a lossy one by smoothing the paths. On Cityscapes at  $1024 \times 512$  resolution, this method (lossless version) reduces the average semantics overhead (class and instance segmentation maps combined) from approximately 0.30 bpp (computed using the rendered maps) to approximately 0.03 bpp.

#### IV. EXPERIMENTS

In this section, we present the experimental results of our codecs on Cityscapes [26] and ADE20k [27], [28] both at  $1024 \times 512$  resolution. The former consists mostly of street scenes and the latter generic everyday images. We test our pipeline using both learned and conventional codecs including JPEG, JPEG 2000, WebP, and BPG as the backbone codec  $c$ . An extensive ablation study will be presented to demonstrate the usefulness of semantics. Note that we focus our evaluations on low and medium bitrates since distortion and perception diverge most severely in this regime and empirically, existing codecs typically fail to provide visually appealing results, which necessitates a better compression method and also helps demonstrate the effectiveness of semantics for improving compression quality.

In these experiments, we keep our instantiation of the general framework simple for two reasons: First, we would

<sup>2</sup> [6] proposed a similar method with limited motivation and details.

like to show that even with a simple implementation and little engineering effort, semantic enhancement can bring significant improvements to perception-accuracy-distortion, which would more effectively prove the usefulness of high-level semantics in compression. Second, this work focuses on evaluating the high-level idea of leveraging semantics instead of pursuing state-of-the-art performance. Thus, we feel that leaving out some sophisticated engineering used in other learned compression works can help simplify the presentation and make the interpretation of results easier.

#### A. On Choice of Baselines

Note that different from most existing works on learned compression, we are not proposing a single new codec. Rather, we are proposing a generic framework that can be used to enhance any given codec. As a result, our experimental evaluations focus on demonstrating the *relative* improvements gained from semantic enhancement with respect to the chosen backbone codecs to show the usefulness of semantics. This contrasts most existing works, where it would make sense to compare one or two newly proposed codecs “in parallel” against existing methods. Therefore, when interpreting our results, each codec should be primarily compared against and only against its semantically-enhanced counterpart.

Also, note that we mostly employ engineered codecs such as BPG instead of learned ones as our backbone codecs due to their speed and memory efficiency. And we extensively test with all popular engineered lossy codecs. At the time of this writing, BPG still performs on par with state-of-the-art learned codecs in terms of R-D [19]. Thus, our results on BPG should let us gauge how much improvement our SE pipeline can bring to advanced learned codecs.

#### B. Experimental Details

1) *Dataset*: For Cityscapes, we trained the models on the densely-annotated training set of 2975 images. For evaluation, we randomly sampled 10 images from each of the three cities in the validation set to create a test set. We used the ground truth class segmentation map and instance boundary map instead of learned ones as the semantics to simplify the set-up and focus more on compression. Nevertheless, we provide results from using learned semantics at test time to demonstrate how our pipeline would perform with the currently available segmentation technology. In these cases, the class (instance) segmentation maps were extracted by a trained DeepLab v3+ [49] (Mask R-CNN [50]). One may also train the compression pipeline with learned semantics or jointly train the segmentation network with the codec and we leave this as a future work.

For ADE20k, we trained the models on the full training set of 20210 images. For evaluation, we randomly sampled one image from each leaf folder in the dataset directory, creating a total of 682 images. We then randomly split this set into a validation set and a test set. During training, we used the best hyperparameters from Cityscapes and saved the best models according to their validation performance on our ADE20k validation set. The images were all reshaped to  $1024 \times 512$ .

Note that the images in ADE20k have already been compressed by JPEG and therefore are no longer suitable for a rigorous quantitative analysis for compression performance. We mainly show qualitative results as proof of concept that our method works well both for specialized images such as street scenes and in the wild for any generic image.

2) *Architecture*: We implemented  $F$  using a generator network with the same architecture as in [43]. The decoded class segmentation map was one-hot encoded to form a  $k$ -channel 3D tensor with height and width the same as the original image, where  $k$  is the number of classes and is 35 for Cityscapes and 151 for ADE20k. To form the input of  $F$ , this segmentation map tensor was channel-concatenated to the decoded RGB image and another single-channel instance segmentation edge map.

For  $c$ , we tested with JPEG, JPEG 2000, WebP, BPG, and an autoencoder-based learned codec. BPG was implemented using the official *libbpg*<sup>3</sup>. JPEG, JPEG 2000, and WebP were implemented with *Pillow*<sup>4</sup>. For the autoencoder, the encoder (decoder) consists of 4 convolutional (transposed convolutional) layers, each downsampling (upsampling) the image by 2. We added a binarization layer in between to quantize the continuous hidden activations to binary codes, using the method specified in [7]. A Prediction by Partial Matching (PPM) [51] coder was used on the bitstream produced by the binarizer to further reduce bitrate. For simplicity, we did not explicitly estimate and minimize the entropy of the bitstream during training.

The semantics was losslessly compressed following descriptions in Section III. A PPM [51] coder was used on the delta encoded boundaries to further reduce bitrate. On the 30-image test set of Cityscapes, the average semantics bpp overhead is 0.03.

3) *Training*: We used LS-GAN [52] as our  $L_{\text{GAN}}$  formulation. For the first and second training phases, we supplemented the GAN loss with VGG loss  $L_{\text{VGG}}$  [34] and GAN feature matching loss  $L_{\text{FM}}$  [43]. For pipelines using conventional codecs as  $c$ , we trained  $F$  with the  $L^1$  loss as the distortion term. For pipelines using learned  $c$ , we trained  $F$  and  $c$  with the  $L^2$  loss to demonstrate that our framework works well with any generic distortion loss. The weightings of the loss terms follow those in [43]. Adam [53] was used as the optimizer and we set the learning rate to 0.0001 with no scheduling. With a batch size of 2, we trained the networks on Cityscapes (ADE20k) for 30, 30, and 50 (5, 5, and 7) epochs for the three phases, respectively. Except for the objective function formulation, all settings remain unchanged across the three phases for a given model. All experiments were performed on an NVIDIA Tesla V100 GPU. The inference time depends on the underlying backbone image codec and its compression settings. As a reference, each test image took on average 1 second to process for JPEG-SE with the backbone JPEG codec running at quality 15<sup>5</sup>.

<sup>3</sup><https://bellard.org/bpg/>

<sup>4</sup><https://github.com/python-pillow/Pillow>

<sup>5</sup>This is using *Pillow*'s metrics. All other JPEG settings were kept at *Pillow*'s defaults for this test.

4) *Evaluation*: To quantify R-P performance, we conducted a comprehensive user study on Amazon Mechanical Turk (MTurk). For each pair of original and SE codecs with comparable bpp values, we created a questionnaire consisting of a sequence of images for 30 distinct, randomly-chosen workers by randomly sampling 10 images out of our 30-image Cityscapes test set, producing a total of 300 responses per comparison. For each image in a questionnaire, we presented the original and reconstructions from the two codecs (baseline codec and its SE counterpart) with relative orders of the latter two randomized. The task for the worker was to choose between the compressed images which one they thought was a better compressed version of the original with the only standard being their subjective opinion. As quality control, three sanity checks were added to random positions in each questionnaire using 3 more randomly chosen test images, in which one of the compressed images was produced by a lossless codec. Failing to choose the lossless codec in any of the sanity checks will cause a response to be rejected and the job reposted to other workers. We also rejected responses finished within 60 seconds to further filter out workers that picked images quickly at random and happened to have passed all three sanity checks by luck.

To measure R-A performance, we used bounding-box object detection<sup>6</sup> as an example to show that the SE codecs produce reconstructions that not only look better to human viewers but also enable machines to achieve better performance on high-level vision tasks. For this study, we used a trained Faster R-CNN [54] from [55] for bounding-box object detection on the codec reconstructions. Detection performance is reported using AP@IoU=.50:.05:.95, the primary challenge metric of COCO 2019 [56].

We used LPIPS [29], MS-SSIM [24], and PSNR as the distortion metrics to quantify R-D performance. LPIPS measures the distance between two images using the deep features learned by a trained classification network. We used the authors’ official implementation (v0.1) with default settings, which uses an AlexNet [57] as the trained network. Note that although LPIPS was named a “perceptual” metric by the authors, it is still a pairwise metric that does not perfectly align with the perception quality [23]. Nevertheless, the authors in [29] showed that LPIPS aligns with human perception much better than the more traditional distortion losses.

### C. Main Results

In this section, we present the main results of our study. Following the trend in learned image compression community [25], we consider human ratings to be the most important metric among perception, accuracy, and distortion and therefore present R-P results first. Quantitative evaluations will be performed on Cityscapes and qualitative results will be provided for both Cityscapes and ADE20k. We refer the readers to the supplementary materials for the qualitative results on ADE20k.

<sup>6</sup>We did not choose segmentation because our pipeline involves transmitting segmentation maps to the receiver’s end. Thus, there would be no need to perform segmentation again.

1) *Rate-Perception*: Fig. 4 quantitatively shows that the SE codecs produce more visually appealing reconstructions compared to the originals. Indeed, the SE codecs are almost always preferred by the human raters, often by a considerable margin. From the visualizations in Fig. 2, it is clear that the SE codecs “understood” the images. To be specific, the original codecs typically failed to respect boundaries between semantic regions, producing image patches that lack realism. This can be seen in between the tires and the road surface, among other places. Indeed, we know that tires are not the same as the road surface in real-world, yet the non-SE codecs do not “understand” this. Even worse, some non-SE codecs produced severe artifacts, which are essentially low-level features that are not possible in real scenes. In comparison, even though the SE codecs could not get all the details correct due to the low bpp allowances, they preserved well the semantic boundaries and they rarely produced low-level features that are incoherent or unnatural.

2) *Robustness Against Noise on Low-Level Visuals*: We have argued and demonstrated that the high-level knowledge gained via training helps the SE codecs produce more natural and perceptually appealing reconstructions. Intuitively, it should therefore also help the SE codecs distinguish between valid low-level features and artifacts for any given semantics. For undesired artifacts such as noise, the SE codecs should be able to filter them out in the reconstruction. As proof of concept, we added Gaussian noise to a test image to be compressed. And from Fig. 5, we can see that the SE codec indeed filtered out the added noise at test time whereas the original codec tried to reconstruct noise as well, which, in most scenarios, is undesirable.

It is worth pointing out that the SE model can achieve this denoising effect without having been trained with noisy examples. In contrast, existing denoising pipelines require noisy images for training [58]–[60]. To obtain noisy images, one either explicitly device a synthetic noise model, which makes the generation process cheap but does not necessarily produce realistic noisy images [59]. Alternatively, one captures real noisy images and sometimes the corresponding ground truth images from the real world, which is certainly a nontrivial procedure [60].

Our work potentially opens up new directions for image denoising research, in which one does not need at all or no longer need as many noisy images for training to achieve the same level of denoising effect. This contrasts a branch of works aiming at removing the need for noisy-clean pairs during training [58] since our framework only requires clean-clean pairs.

3) *Rate-Accuracy*: From Fig. 6, we see that for all of the five codecs, the object detector performs much better on the SE codecs’ outputs. In practice, this means that one can now transmit fewer bits to the receiver’s end for the machine agent to achieve a certain level of performance, that is, the agent can make a better decision faster. For the ever-growing set of machine-based vision tasks including autonomous driving and face recognition, this observation potentially makes the SE codecs more suitable than the existing ones. And considering how rapidly the industry is embracing automated algorithms

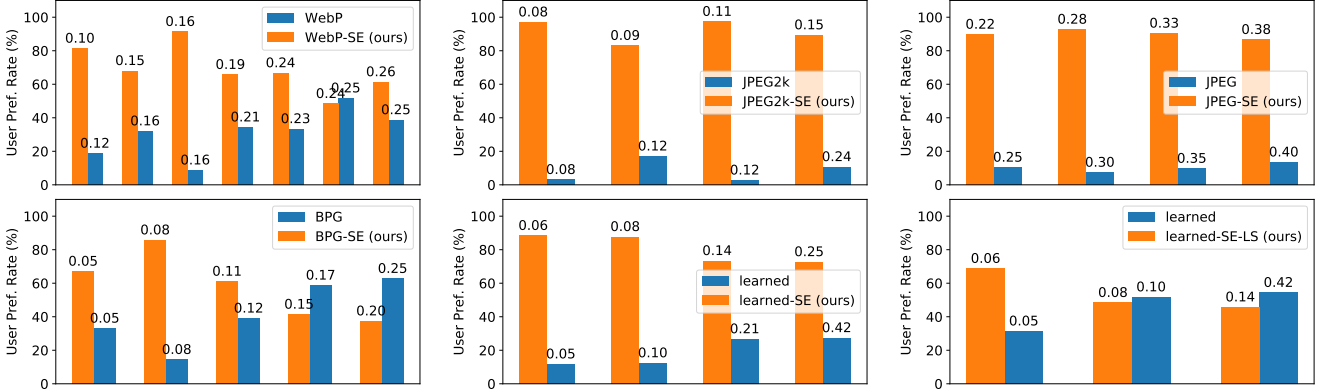


Fig. 4: Rate-perception performance on Cityscapes. Each pair of neighboring bars represent two codecs compared by human raters. The height of each bar shows the % of times this codec is preferred among 300 distinct user responses. Text annotations are the bpp values. The number of bars can differ across subfigures depending on the number of models trained. For this and all results in the paper, the bpp overhead from semantics has *already* been included in the bpp values of the SE codecs. “LS” refers to using learned semantics extracted by off-the-shelf segmentation networks at test time. The SE models are almost always preferred, often by a considerable margin.

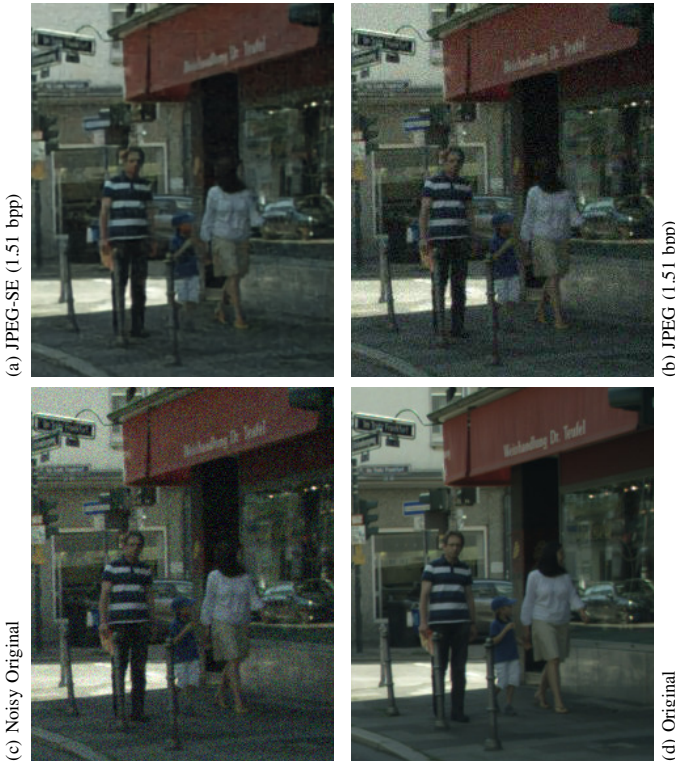


Fig. 5: Their high-level understanding of the scenes robustifies the SE codecs against noise on the low-level visual features without additional training on noisy images.

for vision tasks and how ubiquitous these tasks are, these results are highly relevant and the SE codecs are much better geared toward an AI-powered future. We emphasize that the formulation of our training objective does not directly optimize for rate-accuracy. The improved rate-accuracy performance should be seen as a byproduct from using semantics.

4) *Rate-Distortion*: From Fig. 7, 8, and 2, we see that for all codecs, the SE models achieve comparable or better distortion at most bpp values. On both LPIPS and MS-SSIM, which are the distortion metrics that align more closely with human perception among the three [24], [29], the SE codecs outperform the originals in all cases except for BPG.

Note that to achieve their superior performance in terms of R-P, the SE codecs likely had to compromise R-D to some extent due to the known tension between distortion and perception [22], [23]. Further, since compressing semantics is not the main focus of this work, we always losslessly compressed semantics using the aforementioned simple method, creating a constant bpp overhead of 0.03 that becomes significant in some of the extreme low-bitrate settings in which the evaluated BPG-SE models mainly operated, where it may suffice to use lossy compression instead and reduce rate.

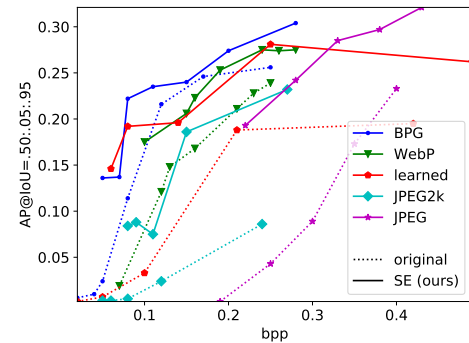


Fig. 6: Rate-accuracy performance on Cityscapes. Each AP value was calculated with the same off-the-shelf bounding-box object detector operating on the codec’s reconstructions at that bpp. The higher the AP, the more accurate the detection. For all codecs, the SE model enables the downstream object detector to detect much more accurately at any given bpp.



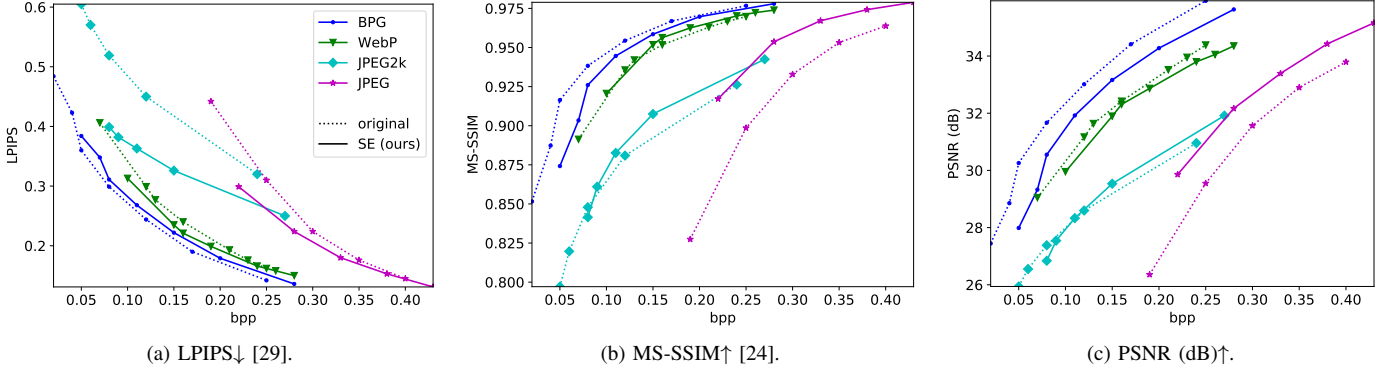


Fig. 7: Rate-distortion performance on Cityscapes.  $\uparrow$  ( $\downarrow$ ) means the higher (lower) the better. Among the three distortion metrics used, PSNR, the most traditional one, has been known to be inconclusive in quantifying compression quality [24], [29]. LPIPS [29] and MS-SSIM [24] have been shown to align much better with human perception. The SE models achieve better or comparable rate-distortion performance, especially when characterized by LPIPS and MS-SSIM.

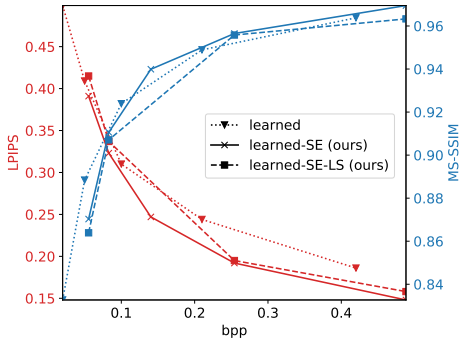


Fig. 8: Rate-distortion performance on Cityscapes. “LS” refers to using learned semantics extracted by off-the-shelf segmentation networks at test time.

Further, as a side note, all SE codecs outperform the originals in nearly all comparisons with the semantics overhead ignored and we refer the readers to the supplementary materials for more details. This represents a performance upper bound of our method that can hopefully be approached with advances in compression techniques tailored for semantics. This, however, is out of the scope of this paper and we hope our work can serve as evidence of the usefulness of semantics and draw more attention to this previously overlooked topic of compressing semantics.

For the models using a learned codec as  $c$ , we also provide results from using learned semantics at test time. These SE models still obtained performance comparable to those using ground truth semantics despite the imperfect segmentation. Naturally, we expect that future advances in semantic segmentation will help close this performance gap.

5) *Comparing With Related Work*: In [6], the authors proposed a version of their GAN-based codec that leveraged semantics (named “selective generative compression” (SC)). Among existing works, this perhaps resembles our work the most. Although [6] only used semantics for bitrate control in a restricted, human-in-the-loop compression use case, their

proposed pipeline is actually similar to ours. In terms of training objective, however, they only optimized for rate-perception performance, whereas we focus on jointly optimizing rate-perception-distortion. Another major difference is that they only focused on one specific codec instantiation, whereas we propose a generic framework that can be used to enhance any codec with semantics.

TABLE I: Comparing against [6]: Average rate-distortion performance on the 3 test images provided by [6].  $\uparrow$  ( $\downarrow$ ) means the higher (lower) the better. Our BPG-SE performs much better than [6] in terms of distortion metrics.

Codec	bpp	LPIPS↓	MS-SSIM↑	PSNR (dB)↑
[6]	0.23	0.203	0.8634	22.50
BPG-SE	0.20	0.196	0.9654	31.56

In Fig. 9, we present qualitative comparisons of one of our codecs (BPG-SE) against theirs. Results from a baseline codec are also illustrated as references. Both [6] and our BPG-SE are able to generate more visually appealing reconstructions than the stock BPG baseline. Indeed, [6] and BPG-SE produce cleaner object boundaries and less compression artifacts. However, due to their disregard for image fidelity, result from [6] can look quite different from original. In contrast, our training objective can simultaneously lead to good perceptual quality as well as small distortion with respect to the original. Table I further quantifies the advantage of our method in terms of image fidelity.

Note that since [6] did not disclose their trained models or code, we could only conduct comparisons with the test images they demonstrated in their paper. With that said, their semantic-aware codec (the SC codec in their paper) can be viewed as a special case of our SE codecs but trained with only Phase 1 and 2. With this connection established, our ablation study on the importance of Phase 3 then serves as another set of large-scale comparisons between our approach and [6]. This ablation study can be found in Sec. IV-D3.



(a) Left to right: [6] (0.18 bpp), original, BPG (0.16 bpp), BPG-SE (0.17 bpp)



(b) Left to right: [6] (0.33 bpp), original, BPG (0.24 bpp), BPG-SE (0.24 bpp)

Fig. 9: Comparing [6] against our BPG-SE. While both [6] and BPG-SE generate more visually pleasing results than the BPG baseline (cleaner object boundaries, less artifacts, etc.), our BPG-SE additionally preserves high image fidelity. Full images and more comparisons are in the supplementary materials.

#### D. Ablation Study

1) *The Importance of Semantics*: To rigorously validate the usefulness of semantics, we performed an extensive ablation study using various codecs. First, we fed dummy all-zero semantics to a trained JP2-SE model during test time. Comparing Fig. 2k with 2j and also the first row of Table II, we conclude that the removal of semantic information at test time significantly worsened performance and in particular, caused the codec to behave similarly to the original JP2 and produce blurred semantic boundaries. However, it may be that the same quality improvement can be achieved using any learned post-processing module without semantics had we removed semantics during training such that the model adapts to its absence at test time. To rule out this possibility, we trained one SE model for each engineered backbone codec from scratch with the semantics channels zeroed out throughout. The GAN loss was left unchanged during training and these models did not use semantics at test time. From Fig. 10b, and 10d, it can be seen from the silhouette of the car that without semantics, the codec returned to producing reconstructions with blurred semantic boundaries and unnatural details. Rows 2-5 of Table II further quantitatively verifies that the use of semantics is indispensable for the improved compression performance of SE codecs.

2) *Contribution From GAN*: The GAN loss in our framework helps improve perception quality since GAN loss can be interpreted as a divergence measure between the original density and that of the reconstruction. On the other hand, GAN is known to create hallucination artifacts even in the presence of a distortion loss [6] and we proposed a distortion-loss-only training phase to mitigate this issue. The question, however, is whether our third training phase washes off all the improvement in perception quality introduced by GAN. To prove that our training scheme enables GAN to contribute to perception quality without introducing severe hallucination artifacts, we trained a JPEG-SE model with the GAN component (hence also semantics) removed throughout and compare it with the same model trained with only the semantics removed. From Fig. 10c and 10d, it is evident that GAN helped produce much more vibrant colors than what could

be achieved with a vanilla learned post-processing module. And from the second and sixth rows of Table II, removing GAN even caused a performance drop in terms of distortion and accuracy. Note that the user preference rates on the two rows were both computed with respect to the SE baseline and therefore cannot be used to conclude that the removal of GAN did not significantly affect the perceptual quality.

3) *The Three-Phase Training Scheme*: Fig. 11b makes it clear that the models trained without Phase 3 produced reconstructions with heavy hallucination artifacts. In comparison, the distortion-loss-only Phase 3 effectively improved the fidelity of the reconstructed images. This, however, came at a price of decreased perception quality and detection accuracy (the last two rows of Table II), which is expected since in the evaluated bitrates, distortion and perception quality diverge severely [23].

#### V. CONCLUSION

In this paper, we proposed a generic framework that can be used to semantically enhance any image codec. Our work is the first to study the efficacy of high-level semantics as a fundamental supplement to low-level visual features in the general compression setting. Moreover, in contrast to how existing works mostly focus only on rate-distortion, we are the first to consider the joint optimization of perception quality and distortion on full-resolution images. We also demonstrate improved accuracy of downstream vision algorithm.

Despite its simplicity, an instantiation of our framework effectively improved the perception-accuracy-distortion performance of existing learned and engineered codecs by enabling them to leverage high-level semantics in the form of segmentation maps. And extensive ablation study was performed to validate that the use of semantics is the key to the improved performance.

#### REFERENCES

- [1] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: multimedia content description interface*. John Wiley & Sons, 2002.
- [2] Google, “A new image format for the web,” <https://developers.google.com/speed/webp>.
- [3] F. Bellard, “Bpg image format,” <https://bellard.org/bpg/>.



Fig. 10: Ablation study: “-sg;tt” means the model has been trained and tested without semantics or GAN; “-s;tt” means the model has been trained and tested without semantics. The effect of semantics cannot be attained with purely low-level visuals even with the help of GAN, as 10b and 10d suggest. 10c and 10d show that GAN can help improve perception quality through creating more vibrant colors without severely increasing distortion or adding hallucination artifacts. Full images are provided in the supplementary materials.

TABLE II: Ablation study: Each table entry corresponds to **bpp/% of times this codec is preferred when compared against the full SE model among 300 distinct user responses/AP@IoU=,50:05:95↑/LPIPS↓/MS-SSIM↑** on the Cityscapes test set. ↑ (↓) means the higher (lower) the better. Notations: “-s” means “no semantics” and “-sg” “no semantics or GAN”. The stage(s) at which the ablation was performed is specified afterwards with “tt” meaning “during both training and test” and “test” meaning “only during test”. “-Phase 3” means the model has been trained without Phase 3. The full SE model has the best overall R-PAD performance in each original-SE-SE (ablated) triplet.

Codec	Original	SE	SE (Ablated Models)	Ablation
				-s; test
JPEG	.12/2.67/.024/.450/.8809	.11/-/.075/.363/.8826	-/13.23/.018/.414/.8416	
JP2	.25/10.33/.043/.310/.8986	.22/-/.193/.299/.9172	-/20.00/.081/.318/.9138	
JP2	.24/10.67/.086/.320/.9263	.15/-/.186/.326/.9075	-/7.33/.069/.364/.9002	-s; tt
BPG	.08/14.33/.114/.299/.9383	.08/-/.222/.311/.9260	-/25.33/.093/.334/.9210	
WebP	.16/8.67/.168/.240/.9517	.16/-/.223/.221/.9563	-/29.67/.213/.238/.9517	
JPEG	.25/10.33/.043/.310/.8986	.22/-/.193/.299/.9172	-/19.67/.073/.332/.9094	-sg; tt
JP2	.08/2.00/.005/.519/.8479	.08/-/.084/.399/.8415	-/96.90/.242/.251/.7989	
WebP	.16/8.67/.168/.240/.9517	.16/-/.223/.221/.9563	-/95.00/.335/.120/.9444	-Phase 3

- [4] D. Taubman and M. Marcellin, *JPEG2000 image compression fundamentals, standards and practice: image compression fundamentals, standards and practice*. Springer Science & Business Media, 2012, vol. 642.
- [5] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. V. Gool, “Soft-to-hard vector quantization for end-to-end learning compressible representations,” in *NIPS*, 2017, pp. 1141–1151.
- [6] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, “Generative adversarial networks for extreme learned image compression,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 221–231.
- [7] G. Toderici, S. M. O’Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthankar, “Variable rate image compression with recurrent neural networks,” in *ICLR (Poster)*, 2016.
- [8] G. Toderici, D. Vincent, N. Johnston, S. J. Hwang, D. Minnen, J. Shor, and M. Covell, “Full resolution image compression with recurrent neural networks,” in *CVPR*. IEEE Computer Society, 2017, pp. 5435–5443.
- [9] J. Ballé, V. Laparra, and E. P. Simoncelli, “End-to-end optimization of nonlinear transform codes for perceptual quality,” in *PCS*. IEEE, 2016, pp. 1–5.
- [10] —, “End-to-end optimized image compression,” in *ICLR*. OpenReview.net, 2017.
- [11] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, “Variational image compression with a scale hyperprior,” in *ICLR (Poster)*. OpenReview.net, 2018.
- [12] O. Rippel and L. D. Bourdev, “Real-time adaptive image compression,” in *ICML*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 2017, pp. 2922–2930.
- [13] D. Minnen, J. Ballé, and G. Toderici, “Joint autoregressive and hierarchical priors for learned image compression,” in *NeurIPS*, 2018, pp. 10794–10803.
- [14] M. Li, W. Zuo, S. Gu, D. Zhao, and D. Zhang, “Learning convolutional networks for content-weighted image compression,” in *CVPR*. IEEE Computer Society, 2018, pp. 3214–3223.
- [15] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. V. Gool, “Conditional probability models for deep image compression,” in *CVPR*. IEEE Computer Society, 2018, pp. 4394–4402.
- [16] N. Johnston, D. Vincent, D. Minnen, M. Covell, S. Singh, T. T. Chinen, S. J. Hwang, J. Shor, and G. Toderici, “Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks,” in *CVPR*. IEEE Computer Society, 2018, pp. 4385–4393.
- [17] M. Tschannen, E. Agustsson, and M. Lucic, “Deep generative models for distribution-preserving lossy compression,” in *NeurIPS*, 2018, pp. 5933–5944.
- [18] L. Theis, W. Shi, A. Cunningham, and F. Huszár, “Lossy image compression with compressive autoencoders,” in *ICLR (Poster)*. OpenReview.net, 2017.
- [19] M. Li, K. Zhang, W. Zuo, R. Timofte, and D. Zhang, “Learning context-based non-local entropy modeling for image compression,” *arXiv preprint arXiv:2005.04661*, 2020.
- [20] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *CVPR*. IEEE Computer Society, 2015, pp. 3156–3164.





Fig. 11: Ablation study: “-Phase 3” means the model has been trained without Phase 3. Without the refinement from the distortion-loss-only Phase 3, the reconstructions suffer heavily from hallucination artifacts and high distortion, indicating that the model has been overly optimizing for perceptual quality. Full images are in the supplementary materials.

- [21] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, “AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks,” in *CVPR*. IEEE Computer Society, 2018, pp. 1316–1324.
- [22] Y. Blau and T. Michaeli, “Rethinking lossy compression: The rate-distortion-perception tradeoff,” in *ICML*, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 2019, pp. 675–685.
- [23] —, “The perception-distortion tradeoff,” in *CVPR*. IEEE Computer Society, 2018, pp. 6228–6237.
- [24] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003, vol. 2. Ieee, 2003, pp. 1398–1402.
- [25] “Workshop and challenge on learned image compression,” 2020. [Online]. Available: <http://challenge.compression.cc/tasks/>
- [26] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *CVPR*. IEEE Computer Society, 2016, pp. 3213–3223.
- [27] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, “Semantic understanding of scenes through the ade20k dataset,” *arXiv preprint arXiv:1608.05442*, 2016.
- [28] —, “Scene parsing through ade20k dataset,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [29] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *CVPR*. IEEE Computer Society, 2018, pp. 586–595.
- [30] “Portable network graphics,” <http://www.libpng.org/pub/png/>.
- [31] “Flif - free lossless image format,” <https://flif.info/index.html>.
- [32] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. A. Storer, “Semantic perceptual image compression using deep convolution networks,” in *DCC*. IEEE, 2017, pp. 250–259.
- [33] M. Akbari, J. Liang, and J. Han, “DSSLIC: deep semantic segmentation-based layered image compression,” in *ICASSP*. IEEE, 2019, pp. 2042–2046.
- [34] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *ECCV (2)*, ser. Lecture Notes in Computer Science, vol. 9906. Springer, 2016, pp. 694–711.
- [35] S. Santurkar, D. M. Budden, and N. Shavit, “Generative compression,” in *PCS*. IEEE, 2018, pp. 258–262.
- [36] H. Zhang, T. Xu, and H. Li, “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks,” in *ICCV*. IEEE Computer Society, 2017, pp. 5908–5916.
- [37] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, “Stackgan++: Realistic image synthesis with stacked generative adversarial networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1947–1962, 2019.
- [38] Z. Zhang, Y. Xie, and L. Yang, “Photographic text-to-image synthesis with a hierarchically-nested adversarial network,” in *CVPR*. IEEE Computer Society, 2018, pp. 6199–6208.
- [39] S. Hong, D. Yang, J. Choi, and H. Lee, “Inferring semantic layout for hierarchical text-to-image synthesis,” in *CVPR*. IEEE Computer Society, 2018, pp. 7986–7994.
- [40] T. Qiao, J. Zhang, D. Xu, and D. Tao, “Mirrorgan: Learning text-to-image generation by redescription,” in *CVPR*. Computer Vision Foundation / IEEE, 2019, pp. 1505–1514.
- [41] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” in *ICML*, ser. JMLR Workshop and Conference Proceedings, vol. 48. JMLR.org, 2016, pp. 1060–1069.
- [42] J. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *ICCV*. IEEE Computer Society, 2017, pp. 2242–2251.
- [43] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional gans,” in *CVPR*. IEEE Computer Society, 2018, pp. 8798–8807.
- [44] T. Park, M. Liu, T. Wang, and J. Zhu, “Semantic image synthesis with spatially-adaptive normalization,” in *CVPR*. Computer Vision Foundation / IEEE, 2019, pp. 2337–2346.
- [45] Q. Chen and V. Koltun, “Photographic image synthesis with cascaded refinement networks,” in *ICCV*. IEEE Computer Society, 2017, pp. 1520–1529.
- [46] X. Qi, Q. Chen, J. Jia, and V. Koltun, “Semi-parametric image synthesis,” in *CVPR*. IEEE Computer Society, 2018, pp. 8808–8816.
- [47] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, “Generative adversarial nets,” in *NIPS*, 2014, pp. 2672–2680.
- [48] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [49] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *ECCV*, 2018.
- [50] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, “Mask R-CNN,” in *ICCV*. IEEE Computer Society, 2017, pp. 2980–2988.
- [51] J. G. Cleary and I. H. Witten, “Data compression using adaptive coding and partial string matching,” *IEEE Trans. Communications*, vol. 32, no. 4, pp. 396–402, 1984.
- [52] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, “Least squares generative adversarial networks,” in *ICCV*. IEEE Computer Society, 2017, pp. 2813–2821.
- [53] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR (Poster)*, 2015.
- [54] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, 2015, pp. 91–99.
- [55] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, “Mmdetection: Open mmlab detection toolbox and benchmark,” *CoRR*, vol. abs/1906.07155, 2019.
- [56] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: common objects in context,” in *ECCV (5)*, ser. Lecture Notes in Computer Science, vol. 8693. Springer, 2014, pp. 740–755.
- [57] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *NIPS*, 2012, pp. 1106–1114.
- [58] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, “Noise2noise: Learning image restoration without clean data,” in *ICML*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 2971–2980.
- [59] B. Mildenhall, J. T. Barron, J. Chen, D. Sharlet, R. Ng, and R. Carroll, “Burst denoising with kernel prediction networks,” in *CVPR*. IEEE Computer Society, 2018, pp. 2502–2510.



- [60] C. Chen, Q. Chen, J. Xu, and V. Koltun, “Learning to see in the dark,” in *CVPR*. IEEE Computer Society, 2018, pp. 3291–3300.

## SUPPLEMENTARY MATERIALS FOR “JPD-SE: HIGH-LEVEL SEMANTICS FOR JOINT PERCEPTION-DISTORTION ENHANCEMENT IN IMAGE COMPRESSION”

### A. FULL TABLE FOR R-D PERFORMANCE

See Table III for the complete results of the R-D performance of all the models that we have trained and tested. We put each SE codec side-by-side against its original to clearly demonstrate how high-level semantics affected codec performance. This provides more insights in addition to the R-D curves in Fig. 7.

In this table, the SE codecs achieve better distortion performance in nearly all comparisons. This comes at the price of an extra 0.03 bpp overhead from semantics which necessarily worsens the R-D performance. Although, note that this overhead is based on our primitive lossless semantics compression method described in Section III. Given that compressing semantics has not been a major research focus in the past and that few works on this topic exist, we believe that more efficient methods can be developed in the future and we hope that our work can draw attention to this largely unexplored area, which would help the SE codecs approach the R-D performance upper bound given in Table III.

### B. FULL IMAGES OF VISUALIZATIONS IN MAIN TEXT

See Fig. 12, 14, 15, and 16 for the full images of Fig. 2, 9, 10, and 11, respectively.

### C. MORE VISUALIZATIONS FOR CITYSCAPES

See Fig. 17, 18, and 19 for more visualizations on Cityscapes. In Fig. 17 and 18, the codecs operate in their respective low bitrate settings whereas for Fig. 19, we demonstrate codecs operating under relatively higher bitrates. In these images, we see the same trend we saw in images from the main text. Specifically, the SE codecs produce results with similar distortion but better perception quality, which can be clearly seen especially in semantic boundaries: No color bleeding is observed from the SE codecs. In contrast, this unnatural and visually unappealing phenomenon is pervasive in the originals when working in the illustrated bitrate settings. This is perhaps because in these low to medium bitrates, the compressed hidden code cannot perfectly preserve all low-level features. In this case, the SE codecs can use their knowledge on what low-level visual features are more likely to appear in each semantic region to make an educated guess on the features given semantics. In contrast, the originals do not have this extra degree of freedom and can only use information in the hidden code, which is highly incomplete and therefore leads to visually unappealing reconstructions.

### D. VISUALIZATIONS FOR ADE20K

See Fig. 20, 21, and 22 for visualizations from ADE20k.

JPEG			JP2		WebP		BPG		learned		learned (LS)	
	Original	SE (Ours)	Original	SE (Ours)	Original	SE (Ours)	Original	SE (Ours)	Original	SE (Ours)	Original	SE (Ours)
bpp	0.19	—	0.05	—	0.07	—	0.02	—	0.02	0.03	0.02	0.03
LPIS	0.442	<b>0.299</b>	0.605	<b>0.399</b>	0.406	<b>0.313</b>	0.484	<b>0.384</b>	0.499	<b>0.391</b>	0.499	<b>0.415</b>
MS-SSIM	0.8274	<b>0.9172</b>	0.7971	<b>0.8415</b>	0.8913	<b>0.9206</b>	0.8516	<b>0.8742</b>	0.8327	<b>0.8703</b>	0.8327	<b>0.8639</b>
PSNR	26.36	<b>29.86</b>	25.94	<b>26.84</b>	29.05	<b>29.96</b>	27.45	<b>27.99</b>	25.11	<b>26.20</b>	25.11	<b>25.88</b>
bpp	0.25	—	0.06	—	0.12	—	0.04	—	0.05	0.05	0.05	0.05
LPIS	0.310	<b>0.224</b>	0.570	<b>0.382</b>	0.299	<b>0.235</b>	0.423	<b>0.348</b>	0.409	<b>0.323</b>	0.409	<b>0.337</b>
MS-SSIM	0.8986	<b>0.9537</b>	0.8197	<b>0.8610</b>	0.9356	<b>0.9518</b>	0.8873	<b>0.9034</b>	0.8884	<b>0.9106</b>	0.8884	<b>0.9071</b>
PSNR	29.55	<b>32.17</b>	26.55	<b>27.54</b>	31.17	<b>31.90</b>	28.85	<b>29.33</b>	26.87	<b>27.84</b>	26.87	<b>27.57</b>
bpp	0.30	—	0.08	—	0.13	—	0.05	—	0.10	0.11	0.21	0.22
LPIS	0.224	<b>0.180</b>	0.519	<b>0.363</b>	0.277	<b>0.221</b>	0.360	<b>0.311</b>	0.310	<b>0.247</b>	0.244	<b>0.195</b>
MS-SSIM	0.9329	<b>0.9671</b>	0.8479	<b>0.8826</b>	0.9420	<b>0.9563</b>	0.9164	<b>0.9260</b>	0.9239	<b>0.9399</b>	0.9489	<b>0.9558</b>
PSNR	31.57	<b>33.39</b>	27.38	<b>28.33</b>	31.64	<b>32.31</b>	30.26	<b>30.55</b>	28.54	<b>29.39</b>	29.80	<b>30.27</b>
bpp	0.35	—	0.12	—	0.16	—	0.08	—	0.21	0.22	0.43	0.46
LPIS	0.176	<b>0.153</b>	0.450	<b>0.326</b>	0.240	<b>0.199</b>	0.299	<b>0.268</b>	0.244	<b>0.192</b>	0.186	<b>0.158</b>
MS-SSIM	0.9532	<b>0.9742</b>	0.8809	<b>0.9075</b>	0.9517	<b>0.9625</b>	0.9383	<b>0.9446</b>	0.9489	<b>0.9565</b>	0.9633	<b>0.9633</b>
PSNR	32.90	<b>34.43</b>	28.60	<b>29.54</b>	32.42	<b>32.87</b>	31.67	<b>31.92</b>	29.80	<b>30.35</b>	30.95	<b>30.49</b>
bpp	0.40	—	0.24	—	0.21	—	0.12	—	0.43	0.46	—	—
LPIS	0.145	<b>0.131</b>	0.320	<b>0.250</b>	0.193	<b>0.166</b>	0.244	<b>0.222</b>	0.186	<b>0.148</b>	—	—
MS-SSIM	0.9638	<b>0.9789</b>	0.9263	<b>0.9425</b>	0.9630	<b>0.9702</b>	0.9544	<b>0.9585</b>	0.9638	<b>0.9695</b>	—	—
PSNR	33.79	<b>35.16</b>	30.96	<b>31.92</b>	33.52	<b>33.80</b>	33.01	<b>33.16</b>	30.95	<b>31.97</b>	—	—
bpp	—	—	0.23	—	0.17	—	0.17	—	—	—	—	—
LPIS	—	—	0.176	<b>0.158</b>	0.190	<b>0.179</b>	0.190	<b>0.179</b>	—	—	—	—
MS-SSIM	—	—	0.9667	<b>0.9723</b>	0.9669	<b>0.9697</b>	0.9669	<b>0.9697</b>	—	—	—	—
PSNR	—	—	33.95	<b>34.05</b>	33.95	<b>34.05</b>	34.41	34.28	—	—	—	—
bpp	—	—	0.25	—	0.25	—	0.25	—	—	—	—	—
LPIS	—	—	0.162	<b>0.150</b>	0.142	<b>0.136</b>	0.142	<b>0.136</b>	—	—	—	—
MS-SSIM	—	—	0.9698	<b>0.9739</b>	0.9767	<b>0.9780</b>	0.9767	<b>0.9780</b>	—	—	—	—
PSNR	—	—	34.38	34.35	35.93	35.63	35.93	35.63	—	—	—	—

TABLE III: Full table for R-D performance of all models. The bpp of an SE codec equals that of its backbone  $c$  plus an extra 0.03 bpp overhead from semantics. The bpp values from the visuals differ for learned and learned-SE because we retrained a new backbone codec  $c$  for the latter. In learned (LS), the SE models used learned semantics extracted by off-the-shelf segmentation networks at test time. The better in each original-SE pair is marked in bold. The SE model almost always achieves superior distortion performance.

(a) JPEG (0.24 bpp)



(c) BPG (0.05 bpp)



(e) WebP (0.11 bpp)



(g) learned (0.21 bpp)



(i) JP2 (0.12 bpp)



(k) Original



(b) JPEG-SE (0.21 bpp)



(d) BPG-SE (0.05 bpp)



(f) WebP-SE (0.10 bpp)



(h) learned-SE (0.14 bpp)



(j) JP2-SE (0.11 bpp)



(l) JP2-SE (-s; test)



Fig. 12: Full images for Fig. 2



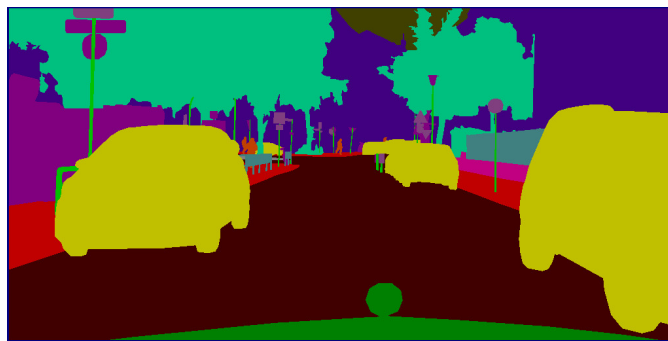


Fig. 13: Semantics (class seg. map) for Fig. 12.

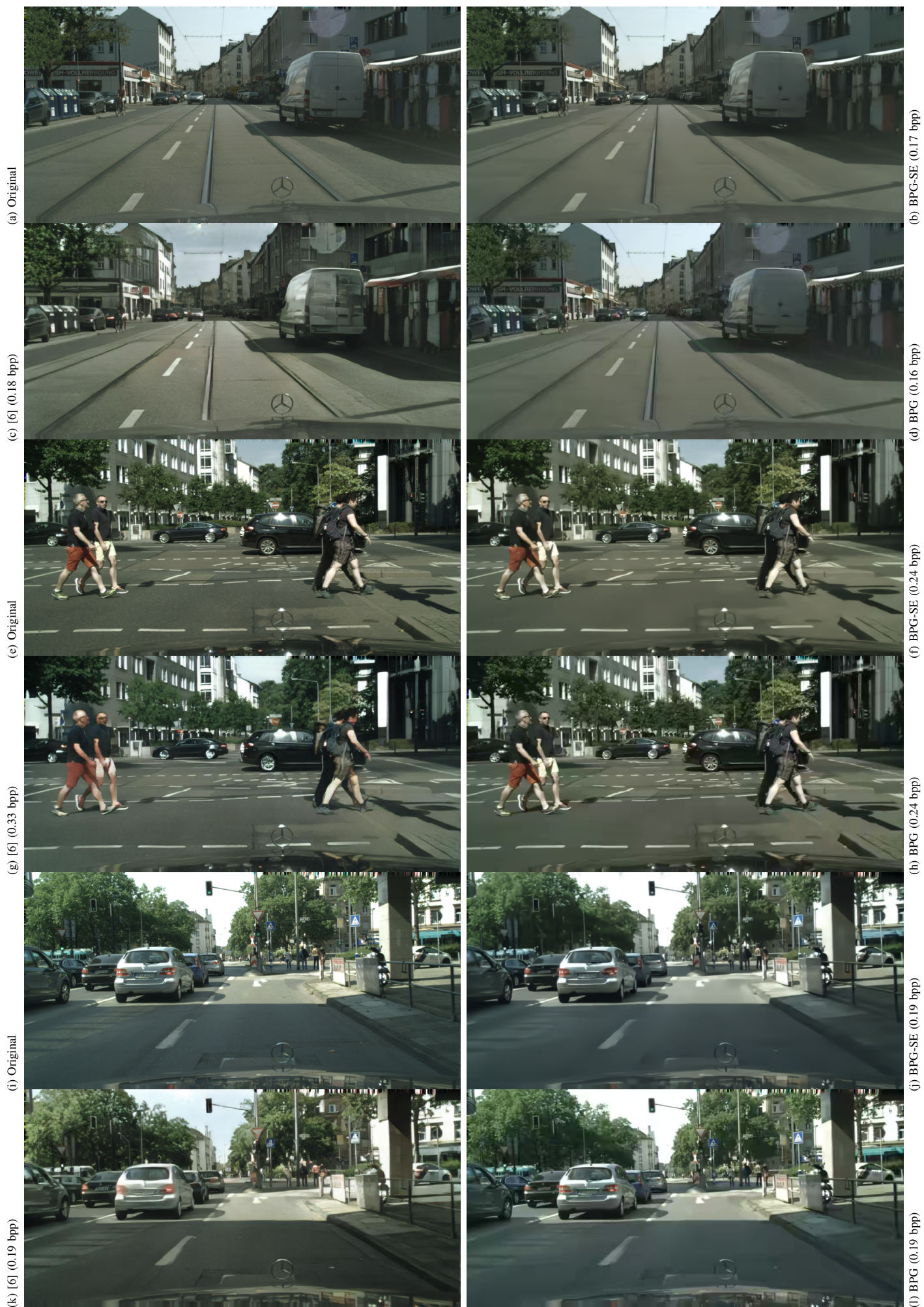


Fig. 14: Full images for Fig. 9 and more comparisons with [6].





Fig. 15: Full images for Fig. 10.

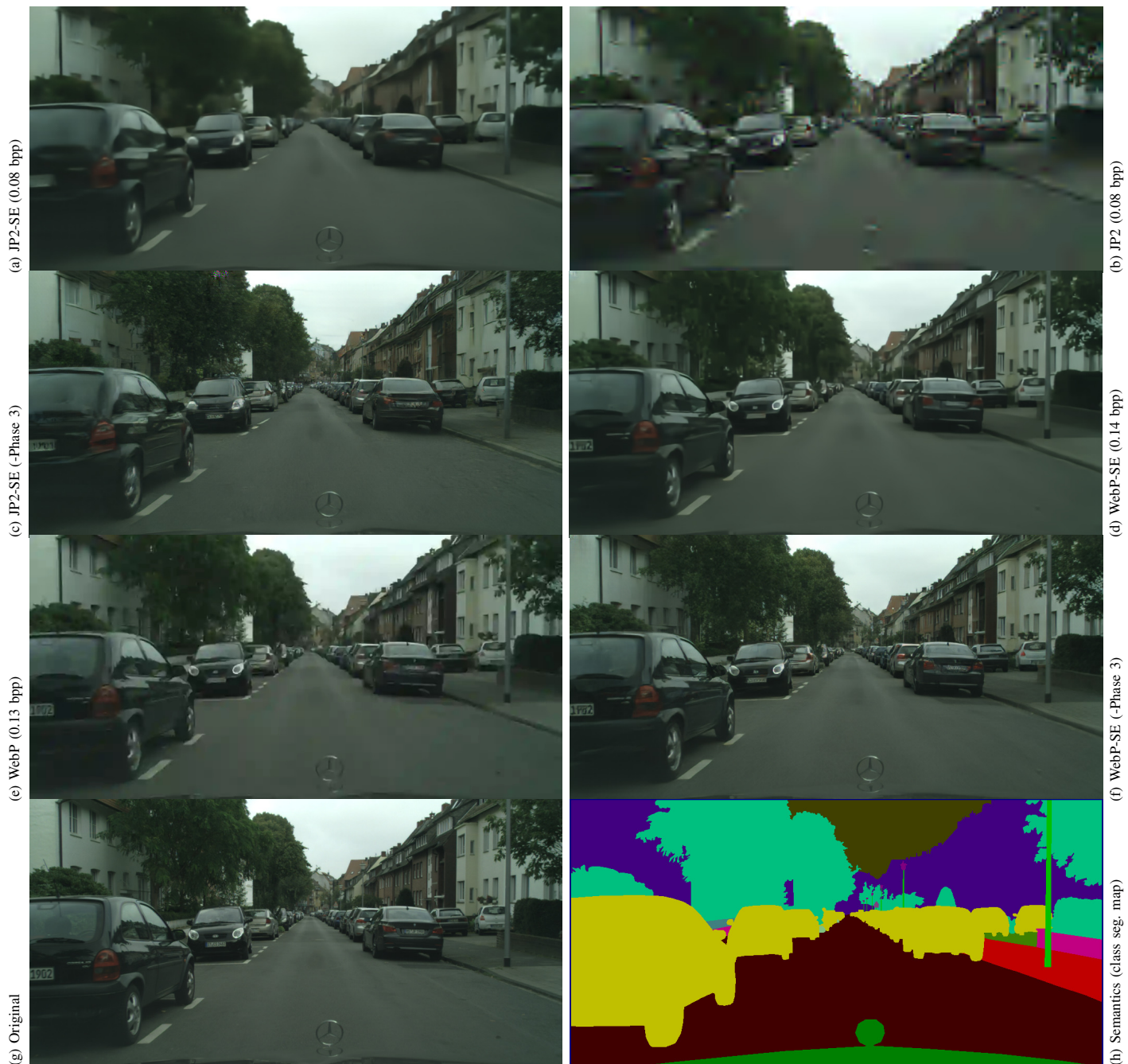


Fig. 16: Full images for Fig. 11. Additionally, we also provide visualization for a WebP model trained without Phase 3, the quantitative performance of which we reported in Table II.



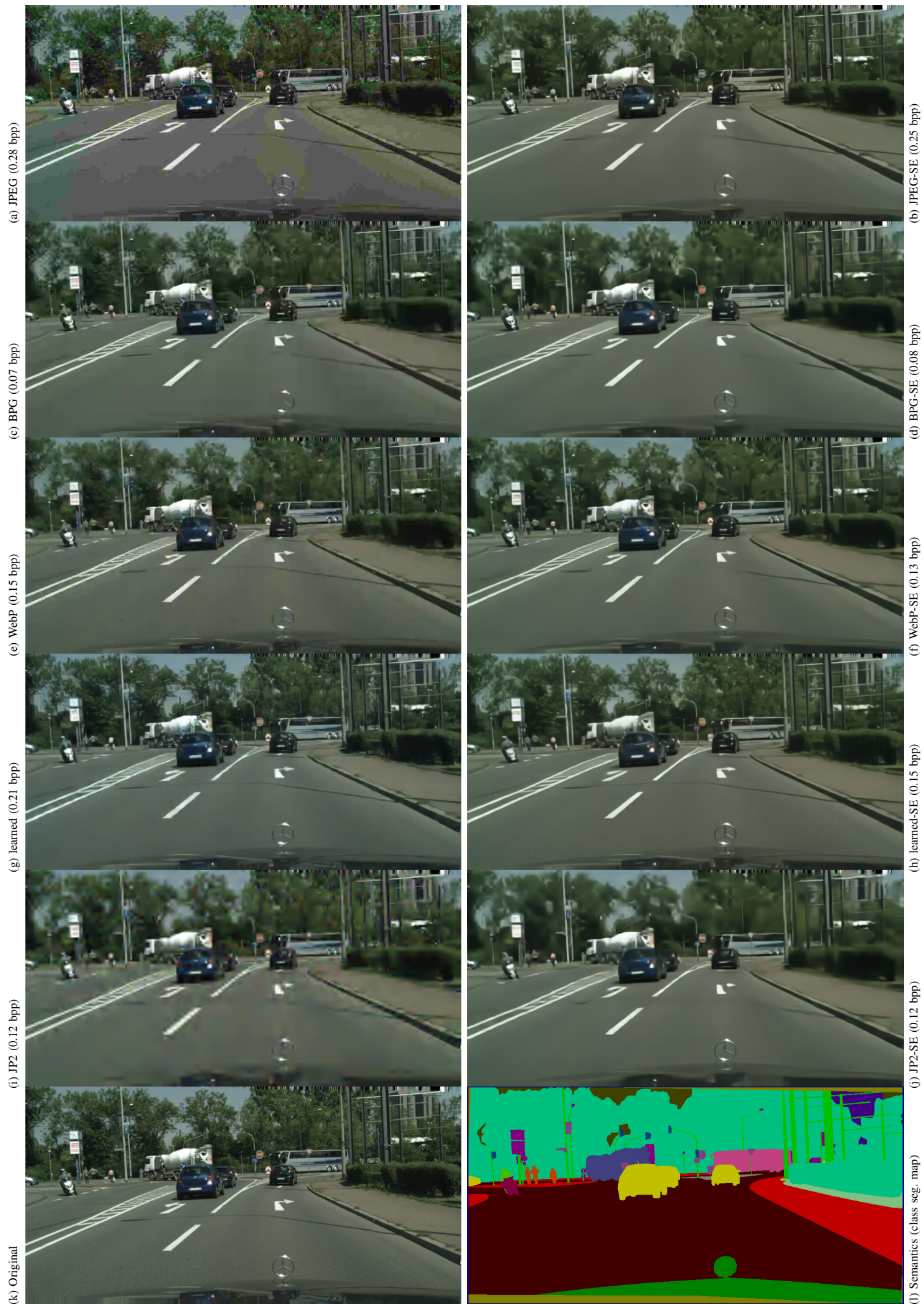


Fig. 17: More visualizations for Cityscapes. These results mainly illustrate codec performance in low bitrate.



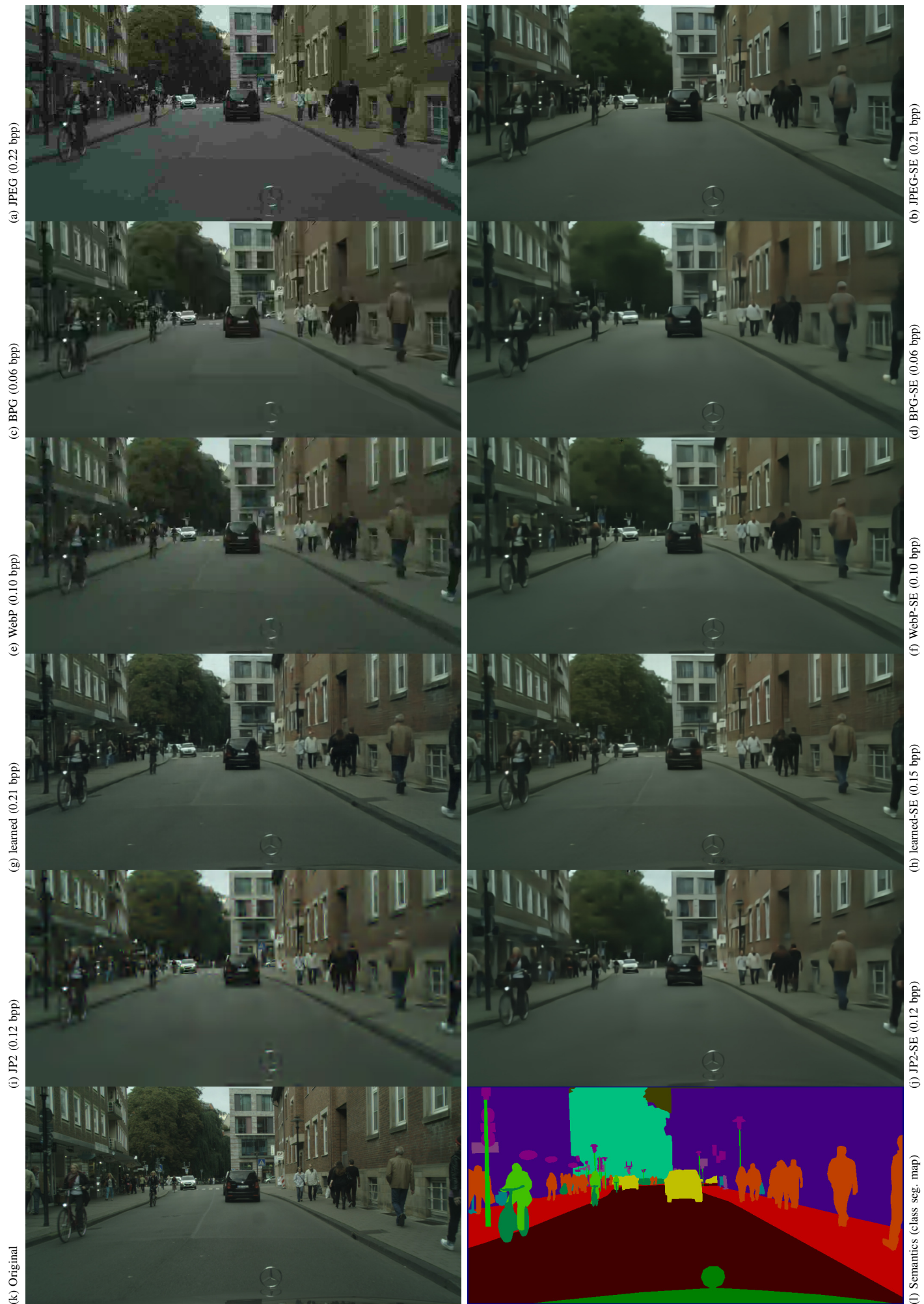


Fig. 18: More visualizations for Cityscapes. These results mainly illustrate codec performance in low bitrate.



(a) JPEG (0.35 bpp)



(c) BPG (0.11 bpp)



(e) WebP (0.20 bpp)



(g) learned (0.21 bpp)



(i) JP2 (0.24 bpp)



(k) Original



(b) JPEG-SE (0.32 bpp)



(d) BPG-SE (0.11 bpp)



(f) WebP-SE (0.19 bpp)



(h) learned-SE (0.15 bpp)



(j) JP2-SE (0.16 bpp)



(l) Semantics (class seg. map)

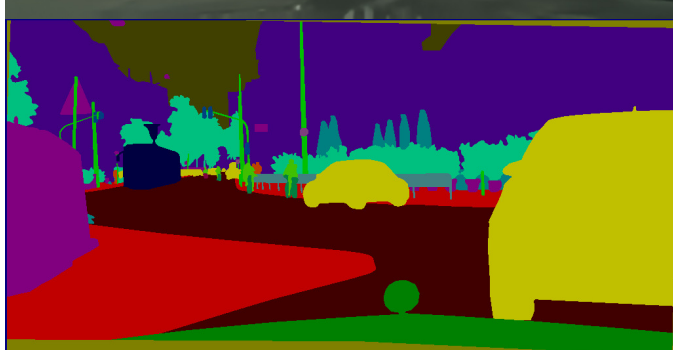


Fig. 19: More visualizations for Cityscapes. In contrast to Fig. 12, 17, and 18, here we mainly demonstrate codec performance in medium bitrate, in which the perceptual quality difference is still visible.



(a) JPEG (0.31 bpp)



(c) BPG (0.09 bpp)



(e) WebP (0.17 bpp)



(g) learned (0.09 bpp)



(i) JP2 (0.06 bpp)



(k) Original



(b) JPEG-SE (0.24 bpp)



(d) BPG-SE (0.06 bpp)



(f) WebP-SE (0.13 bpp)



(h) learned-SE (0.07 bpp)



(j) JP2-SE (0.07 bpp)



(l) Semantics (class seg. map)



Fig. 20: Visualizations for ADE20k. Note how the original image is clearly noisy but all the SE codecs filtered out the noise whereas some originals tried to reconstruct noise as well when given enough bpp allowance (e.g., WebP and JPEG).





Fig. 21: Visualizations for ADE20k.



(a) JPEG (0.33 bpp)



(c) BPG (0.11 bpp)



(e) WebP (0.21 bpp)



(g) learned (0.11 bpp)



(i) JP2 (0.06 bpp)



(k) Original



(b) JPEG-SE (0.27 bpp)



(d) BPG-SE (0.08 bpp)



(f) WebP-SE (0.17 bpp)



(h) learned-SE (0.10 bpp)



(j) JP2-SE (0.09 bpp)



(l) Semantics (class seg. map)

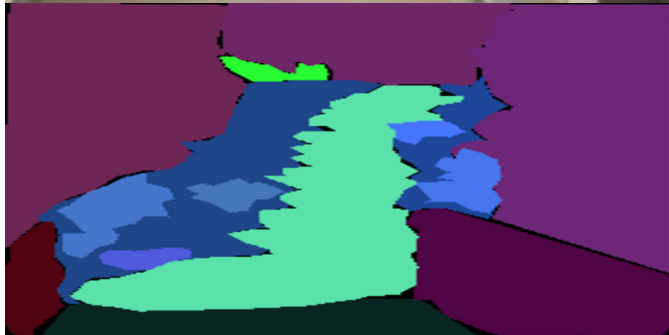


Fig. 22: Visualizations for ADE20k.