A Novel Approach to Evaluating Evaluation Metrics for Multi-Output Structured Prediction

Akshay Vyas, Angelo Pimienta, Nicholas Ruozzi

The University of Texas at Dallas
akshay.vyas, angelomatthew.pimienta, nicholas.ruozzi@utdallas.edu

Abstract

In multi-output structured prediction tasks, while only one ground truth label may be provided in the training data, multiple equally valid outputs may be possible, making reliable evaluation a persistent challenge. We postulate that human evaluators implicitly use task-specific invariants, e.g., object boundaries in colorized images or named entities in translations, to judge if an output is acceptable. Under this assumption, we introduce a notion of approximate task-specific invariants and use them as diagnostic tools to evaluate a variety of existing metrics for vision and language tasks. We use these task invariants as part of a framework to systematically test metric reliability by encouraging domain-relevant invariants in model outputs via an augmented loss function. In our experiments, we observe that enforcing invariants with an augmented loss yields substantial improvements in popular distributional metrics while more traditional metrics change only marginally. Through this invariants-driven evaluation, we expose where standard metrics fail to detect meaningful differences, and we highlight the conditions under which distributional metrics succeed or still fall short.

Code —
https://github.com/akshay25vyas/invariants-metric-eval

Introduction

Machine learning models trained for multi-output vision and language tasks often face an evaluation paradox: there may exist many plausible outputs for a given input, e.g., different colorizations of a photo or multiple valid translations of a sentence, yet traditional evaluation metrics assume a single ground-truth output. In vision, pixel-level metrics such as mean squared error (MSE) or peak signal-to-noise ratio (PSNR) treat any deviation from the single ground-truth image as an error, even if the deviation is a valid alternative, e.g., a shirt colored blue instead of green. These metrics can also be poor proxies for perceptual quality, e.g., an image can have low MSE yet look unnatural to humans. In language, n-gram overlap metrics such as Bilingual Evaluation Understudy Score (BLEU), while standard, can be brittle: a translation that uses synonyms or different phrasing will get a low BLEU score despite conveying the same meaning.

Overall, when trained models output one of many acceptable answers, the above metrics can often over penalize legitimate variation while under penalizing more human perceptible errors, which can lead trained models to produce labels that may have low error with respect to the chosen metric but be easily identified as poor by human evaluators.

Recently, distributional/learned metrics have emerged as attempts to address the above issues. Distributional metrics compare the distribution of labels produced by a given model to the distribution of the ground truth labels. For images, the Fréchet Inception Distance (FID) (Heusel et al. 2017) computes the Wasserstein-2 distance between multivariate normal distributions fit to model generated labels and ground-truth labels in a feature space extracted from the Inception network. As only distributions are considered, these metrics allow for diversity beyond pixel agreement. For text, BERTScore leverages pre-trained language model embeddings to measure semantic similarity between the model-generated text and ground-truth reference. Distributional metrics can better handle paraphrasing or varying image styles, allowing them to be less influenced by superficial differences between the predicted and the ground-truth labels. However, an open question remains: Do these distributional metrics capture the variations that human evaluators expect, or do they have blind spots of their own? For example, if a translation preserves meaning but shifts tone, will Fréchet BERT Distance (FBD) register the change? Does FID drop when an image's colors violate object boundaries, or could it be fooled by overall color statistics? These questions highlight the need for a systematic study of metric sensitivity to meaningful variations in output.

The main postulate in this work is that human evaluators implicitly or explicitly apply domain or task invariants when judging neural network outputs. Invariants are properties that should remain (approximately) unchanged under valid transformations of the output. For example, a human evaluator for the image colorization task might expect that the addition of color does not alter the image's structural details, e.g., the model colorized image should respect the same object boundaries and depth cues as the original ground-truth color image. If neural network colorization causes "color bleeding" across object edges, humans might immediately flag the colorization as incorrect, even if the colors appear plausible overall. Similarly, when assessing machine trans-

lation, bilingual speakers might expect the translation to preserve named entities and sentiment, e.g., a translated sentence that flips the sentiment or mistranslates a proper name violates one of these invariants. In image captioning, an accurate caption should reflect the image's content and relationships, e.g., who is doing what to whom. So, properties like semantic roles and scene relationships act as invariants, e.g., if the image shows a cat on a couch, predicted captions should maintain that subject-object relation.

While the above invariants are natural for their respective domains, many existing metrics do not explicitly account for these kinds of invariants. Further, existing work has observed that neural networks trained for structured prediction tasks do not necessarily produce predictions that enforce the task constraints, e.g., neural networks trained to predict shortest paths need not produce a connected path, a natural invariant for path problems (Ahmed et al. 2022). We observe the same difficulties for the multi-output structured prediction tasks described above: task invariants are not automatically preserved through neural network training (even when training with large data sets). However, as violation of these invariants can result in poor-quality labels that are readily identified as such by human evaluators, our aim is to understand to what extent natural task invariants are captured by existing metrics. We propose a framework to evaluate the evaluation metrics themselves using approximate invariants as a probe. Rather than artificially introducing invariant violations, we systematically compare outputs from standard baseline models (which can significantly violate the task-specific invariants) with invariant-regularized models explicitly trained to respect approximate invariants, e.g., segmentation consistency in image colorization, named entities in translation, etc. We consider three modeling tasks: image colorization, machine translation, and image captioning. For each task, we first define one or more approximate task invariants (see Table 1 for a list of possible invariants). We then train models that are both invariant aware, where we enforce the invariants with an auxiliary loss function, and invariant agnostic. We evaluate the different models via a battery of metrics including standard metrics and distributional metrics and observe how the performance changes when the invariant is or is not enforced. Our contributions are as follows.

- We introduce a novel evaluation framework that uses approximate task invariants as a diagnostic lens for evaluating metrics that provides an alternative to human evaluators.
- We apply our framework across vision and language domains (image colorization, image captioning, and machine translation) with distinct invariants and evaluation challenges. This comprehensive study demonstrates the versatility of our approach and provides insights into metrics ranging from pixel-level (MSE) and token-level (BLEU) to distributional (FID) and embedding-based (BERTScore).
- Our experiments show that distributional metrics, e.g., FID, FBD, etc., are generally more sensitive to invariant-related differences, and enforcing invariants in outputs can lead to significant performance gains ($15\% - 30\%$) in sensitive metrics. In constrast, we find that traditional metrics, e.g., BLEU, PSNR, etc., are largely unchanged when training with the selected invariant losses.

Main Task	**Invariant Task(s)**
Image Colorization	Image Segmentation, Depth Estimation, Edge Detection
Machine Translation (NMT)	Named Entity Recognition (NER), Sentiment Analysis, Paraphrase Detection
Image Captioning	Semantic Role Labeling (SRL), Scene Graph Consistency, Named Entity Recognition (NER)
Text Summarization	Named Entity Recognition (NER), Sentiment Analysis, Coreference Resolution
Video Prediction	Optical Flow Estimation, Object Tracking, Action Recognition
Natural Language Inference	Paraphrase detection, Sentiment Analysis

Table 1: Tasks and corresponding invariants. BOLD tasks are considered in this work.

Related Work

While the aim in this work is to provide alternatives to human evaluators, one branch of related work focuses on metrics that better align with human judgment by incorporating semantic information or learning directly from human evaluations. METEOR (Banerjee and Lavie 2005) extended BLEU by accounting for synonyms and stemming, CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015) reweights n-grams by TF-IDF to reflect their salience, and SPICE (Anderson et al. 2016) incorporates semantic propositional knowledge. SPICE was shown to achieve a higher correlation with human caption preferences than BLEU or CIDEr (Anderson et al. 2016). Hasler and Suesstrunk (2003) introduced Colorfulness to measure vibrancy in images.

Learned metrics based on neural network embeddings have also gained traction. BERTScore (Zhang et al. 2020) leverages pretrained language model embeddings to judge text similarity, yielding significantly improved correlation with human assessments of translation quality. In vision, the Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018) metric measures image differences in a deep feature space. These learned or "feature-based" metrics effectively compute distances in a feature space extracted from trained models on large datasets, making them sensitive to high-level features.

Distributional metrics evaluate the distribution of outputs as a whole, rather than scoring individual output-reference pairs. This approach arose prominently in image generation research. For example, FID (Heusel et al. 2017) has become a popular approach for assessing the realism of images produced by GANs. For language tasks, Xiang et al. (2021) introduce Fréchet Bert Distance (FBD)

and argue that it better aligns high human judgment than BLEU, METEOR, and ROUGE-L.

Distributional metrics also have important limitations. Chong and Forsyth (2020) showed that FID is a biased es-

timator and proposed a corrective fix. Parmar, Zhang, and Zhu (2022) showed that compression and resizing have a significant effect on FID. Jayasumana et al. (2024) show that FID is sensitive to complex image distortions and does not behave monotonically as the distortions increase. They propose using maximum mean discrepancy (MMD) calculated over CLIP embeddings instead as MMD is unbiased, CLIP embeddings are richer than Inception features, and MMD better aligns with human evaluators in their experiments.

Problem or domain invariants have historically been used to improve generalization, e.g., data augmentation for rotation equivariance in computer vision tasks. Some neural architectures are built on invariants, e.g., convolutional layers enforce translation equivariance symmetry, and further research has aimed at hard-coding other symmetries, such as rotation equivariance, into convolutional neural networks (Cohen and Welling 2016; Worrall et al. 2017). However, to the best of our knowledge, the more complex, approximate invariants, including the types we formally define herein, have not been treated in any systematic way.

Another related line of work has explored enforcing certain desired properties through multi-task learning or auxiliary losses. In image colorization, researchers have added auxiliary segmentation or depth prediction tasks to the model so that the colorization network jointly learns to respect object boundaries or 3D structure. Similarly, in machine translation, one might regularize a model to preserve named entities or sentiment by jointly training on a related classification task. These auxiliary loss approaches, e.g., (Misra et al. 2016; Ruder 2017), can act as a form of regularization. However, the approaches in prior work require extra supervision, e.g., segmentation labels, or complex multi-task training setups. Contrast this with the approach taken here in which no additional labels are needed – only access to a pretrained DNN for a desired invariant is required. There has also been progress in explicitly modeling the multi-modal nature of outputs, e.g., using probabilistic frameworks or adversarial generative models that can produce multiple different valid outputs (Zhu et al. 2017; Zhao, Ma, and Ermon 2018). These methods address the inherent ambiguity by aiming to capture the full distribution of outputs, but they often rely on adversarial training and do not guarantee that specific semantic properties are preserved in each output.

Multi-Output Tasks and Invariants

We consider a variant of the standard supervised structured learning task: for data observations $x^{(1)}, \ldots, x^{(M)} \in \mathcal{X}$ our aim is to produce corresponding labels $y^{(1)}, \ldots, y^{(M)}$ that are assumed to be taken from a (structured) set $\mathcal{Y}$. However, we do not assume that there is a one-to-one mapping between the data observations and the labels. Specifically, we assume that there exists a function $f : \mathcal{X} \rightarrow 2^{\mathcal{Y}}$ such that for all $m \in \{1, \ldots, M\}$, $y^{(m)} \in f(x^{(m)})$. Note that if f maps each element of $\mathcal{X}$ to a unique label, i.e., $|f(x)| = 1$, then this is equivalent to a standard supervised learning task whose aim is to find f, but more generally, there could exist $x \in \mathcal{X}$ with $|f(x)| > 1$. In the general case, our aim will be to produce **one** valid labeling for each given data observation (note that one could also consider fitting distributions over possible labels, but we do not consider that task here).

As an example, consider the image colorization task which seeks to transform grayscale images into colored variants. Inherently, the task is ill-posed: a single grayscale image can be mapped to multiple, plausible colorized images. In this task, images are often represented in CIELAB color space as a triple (L, a, b). To map this problem into the above framework, we can choose $\mathcal{X}$ to be a set of grayscale images in Lab color space, $\mathcal{Y}$ to be a set of color images in Lab color space, and f can be defined as the function that maps grayscale images to a set of valid/realistic colorizations. Similarly, consider the task of captioning images: We can define $\mathcal{X}$ as the set of all uncaptioned real-world images, $\mathcal{Y}$ to be a set of captions, and f can be defined as the function that takes an uncaptioned image to a set of valid captions.

In both of the above examples, there is not necessarily a unique label $y \in \mathcal{Y}$ for every $x \in \mathcal{X}$, nor is it clear that there is even a notion of a "best" label $y \in \mathcal{Y}$ as "best" in both of these instances could be subjective. This creates difficulty when using standard approaches to learning/evaluation in these settings: for a given $x \in \mathcal{X}$, there could be $y, y' \in \mathcal{Y}$ with $y \neq y'$ that are both "good" label choices, but they could be far apart from each other under typical notions of distance, e.g., mean squared error, on $\mathcal{Y}$.

Approximate Task Invariants

Consider the image colorization task for which a DNN is trained to produce a colored image in Lab space given a grayscale image. This DNN, when applied to an unseen test image, say of a pedestrian crossing the street, may actually produce a "good" colorization, but that colorization may be evaluated poorly under a squared error metric if the colors deviate significantly from the ground truth, e.g., the pedestrian's shirt is colored red by the DNN in contrast to blue in the ground truth. Different colorizers may produce different colors for the pedestrian's shirt, and unless the pedestrian's shirt is somehow unique, even a human evaluator might not know which color is correct. However, if a particular colorizer assigns a color to the pedestrian's shirt that then bleeds into the neighboring pixels, then a human evaluator may observe that this colorizer has made a mistake.

In the above example, the human evaluator only noticed the error because an unexpected artifact was observed. We can view this as a segmentation error - colors should not spill across object boundaries, or put another way, we expect the segmentations of the ground truth and the colorizer-produced image to (approximately) match. We call this observation that the outputs of "good" colorizers should have the same segmentations as the ground-truth images an *approximate task invariant* for the colorization problem. Formally, we will define invariants as follows.

Definition 0.1. Let $g : \mathcal{Y} \rightarrow \mathcal{Z}$. A labeling problem, as defined above, is **g-invariant** if for all $x \in \mathcal{X}$ and all $y, y' \in f(x)$, $g(y) = g(y')$, i.e., for all data observations x, g applied to any valid labeling of the data observation x yields the same result.

We will typically consider $\mathcal{Z}$ to be a subset of $\mathbb{R}^n$ for some

n. Ideally, $\mathcal{Z}$ should be a space on which similarities/differences are easier to detect than on $\mathcal{Y}$. From the definition of g-invariant, we have that if $||g(y) - g(y')|| > 0$, then there does not exist an $x \in \mathcal{X}$ such that $y, y' \in f(x)$, i.e., we can use $||g(y) - g(y')||$ as a surrogate measure for how close y' is to being in the set $f(x)$. Not all invariants, as defined above, are useful for classification: constant functions, e.g., $g(y) = 0$, trivially satisfy this definition. Note also that the definition makes no restrictions on $y \notin f(x)$. That is, the invariant is a property of all "good" labels, but some labels that are not "good" can map to the same element of $\mathcal{Z}$. As a result, these invariants should be thought of as desirable properties for "good" labelings, but they place no restrictions on invalid labels. However, the invariants that will be most useful in practice should distinguish "good" labels from "poor" labels as much as possible.

The intuition behind considering invariants of this form is that they can act like regularizers. That is, suppose we trained a DNN model for a labeling task, and on input x with ground truth label $y \in f(x)$ it produces the label y'. If $y \neq y'$, the output DNN should still be considered good if $y' \in f(x)$ or if y' is at least close to some member of $f(x)$. If the labeling task is g-invariant, then we can add an auxiliary loss to the objective to penalize model weights for which $||g(y) - g(y')||$ is large. This has two benefits: (1) it allows some deviation from $||y - y'||$ in order to better match the specified invariant, and (2) even if $||y - y'||$ is small, the penalty encourages the model to produce labels that approximately preserve the invariant, which hopefully increases the realism of the resulting predictions.

While true invariants might be difficult to obtain, we can still apply this approach with approximate invariants:

Definition 0.2. Let $g : \mathcal{Y} \to \mathcal{Z}$. A labeling problem, as defined above, is an **ϵ-approximate g-invariant** for some $\epsilon > 0$ if for all $x \in \mathcal{X}$ and all $y, y' \in f(x)$, $||g(y) - g(y')|| \leq \epsilon$ for an appropriately chosen norm on $\mathcal{Z}$, i.e., for all data observations x, g applied to any valid labeling of the data observation x yields *almost* the same result.

Consider again the problem of image colorization. A reasonable (approximate) invariant for this task is that "good" colorizations should be almost indistinguishable to a well-trained image segmentor. Similarly, good colorizations of the same gray scale images should have similar depth maps. In the invariant framework above, a DNN trained for image segmentation or depth estimation, while not a perfect invariant for the colorization task as neither is likely to be error free, can be used to define a function g such that the colorization task is approximately g-invariant. This is the approach that we take in this work.

Invariants as Regularization

Existing work has observed that enforcing task constraints can yield noticeable performance improvements in practice (Ahmed et al. 2022). Motivated by this observation, our aim, then, is (1) to understand to what extent natural task invariants are preserved by/reflected in existing metrics and (2) to determine whether or not enforcing task invariants has any performance benefits for multi-output structured prediction tasks. To encourage models to better align with invariants, we take a common approach: we add an additional penalty to the loss function.

Assume that we are fitting a DNN, represented by f'_θ, which takes as input a data observation x and attempts to produce the ground truth label y. We also assume that we are given an approximate invariant g for this labeling task. With this notation, our approach implements an augmented loss function, L, given by

$$L(\theta|x, y) = L_T(f'_\theta(x), y) + \alpha L_I(g(f'_\theta(x)), g(y)),$$

where L_T is a task-specific loss measuring the error between the DNN's prediction and the ground truth label y, θ is the vector of parameters for the DNN, L_I is the invariant loss for invariant g, measuring the error between the invariant map applied to the ground truth label y and the invariant map applied to the output of the DNN, and α is a hyperparameter controlling the trade-off between these two losses. Note that the invariant g is treated as **fixed**, and although it will be chosen to be a DNN in our formulation, its parameters are frozen, i.e., they are not learned as part of the framework. As an example, for image colorization with a segmentation invariant, we select g to be a frozen semantic segmentation network and L_I to be the cross-entropy between the segmentation masks of the generated and ground-truth color images.

Experimental Evaluation

The aim of training with invariants is to encourage the trained model to produce predictions that, while they may not perfectly match the ground truth, deviate from the ground truth in ways that are approximately preserved by the invariant. The question that we seek to answer in this section, then, is to what extent invariance preservation is reflected in existing metrics. In particular, we are interested in understanding (1) to what extent each metric is sensitive to differences with respect to the selected invariants and (2) to what degree enforcing invariants yields performance improvements in practice.

We illustrate our framework by applying it to the image colorization, machine translation, and image captioning tasks. All models were trained from scratch to ensure that the same training process was used for both the baseline models and the baseline+invariant models. Note that in all experimental settings the weights of the invariant networks are frozen during training. Full implementation details for the baselines can be found in the original papers, and all architectures were obtained directly from the authors of these works. All systems were trained using either one NVIDIA A100 GPU or one V100 GPU. In Table 2, we summarize the choice of baselines, invariants, data sets, losses, and metrics for each experimental setting.

Sensitivity of Metrics to Invariants

Ideally, we would want a model for the selected tasks that produces labels that align well with the ground truth, both in terms of the loss used for training and the chosen invariant loss. However, different evaluation metrics may be more or less sensitive to changes with respect the invariant loss

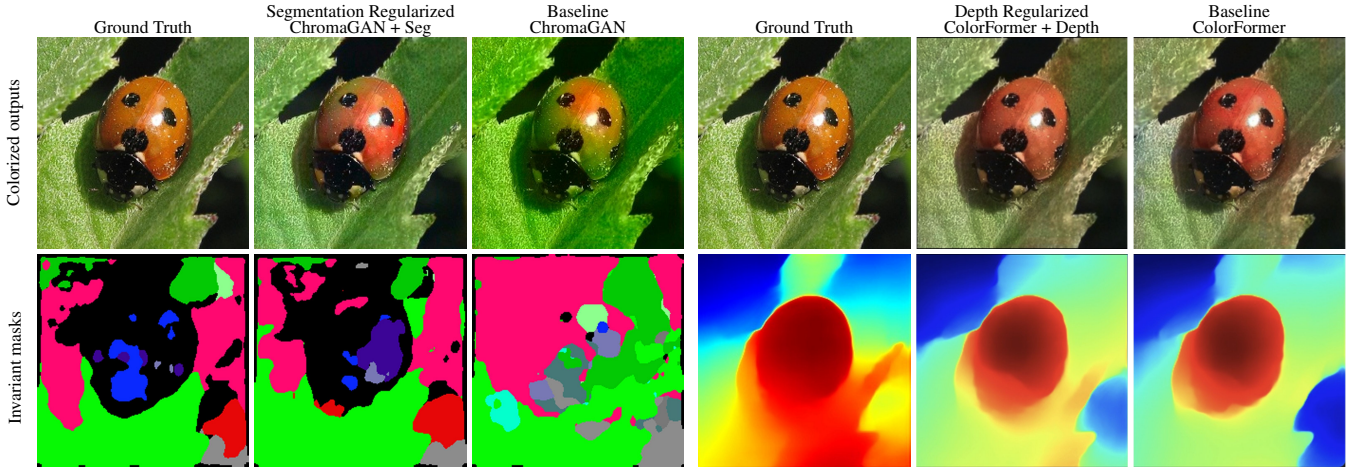


Figure 1: Effectiveness of invariant-based regularization in improving image colorization. Left panel: Comparison using segmentation invariant. Right panel: Comparison using depth invariant. Each panel shows the colorized outputs (top row) and corresponding invariant masks (bottom row). Invariant-regularized models significantly preserve critical structural details, such as object boundaries and spatial depth cues, compared to baseline models that exhibit noticeable artifacts and inaccuracies.

Task	Image Colorization	German to English Neural Machine Translation	Image Captioning
Base Networks	ChromaGAN (Vitoria, Raad, and Ballester 2020), ColorFormer (Ji et al. 2022), and DDColor (Kang et al. 2023)	BiBERT (Xu, Van Durme, and Murray 2021) and Bi-SumCut (Gao et al. 2022)	SmallCap (Ramos et al. 2023) and ClipCap (Mokady, Hertz, and Bermano 2021)
Loss Function (L_T)	Pixel-reconstruction MSE, L1-loss, Perceptual-loss, GAN-loss, Colorfulness-loss	Label Smoothed Cross Entropy: Cross-entropy loss measuring predicted vs target word probability differences with label smoothing for better generalization	Cross Entropy Loss: Token-level classification loss between predicted and ground truth caption sequences
Invariant Networks	Depth: MiDaS (Ranftl et al. 2020) Segmentation: Side Adapter Network (SAN) (Xu et al. 2023)	NER: Hugging Face's (Wolf et al. 2019) Sentiment: Hugging Face's distilbert-base-uncased-finetuned-sst-2-english	NER: Hugging Face's (Wolf et al. 2019) Scenegraph: sngparser (Wu et al. 2019) (Schuster et al. 2015)
Invariant Loss (L_I)	SegmentationSANLoss: Dice loss between predicted vs GT SAN segmentation masks prevents color bleeding DepthMidasLoss: SILog loss between predicted vs GT MiDaS depth maps preserves spatial structure	NER Loss: KL divergence between predicted vs GT BERT-NER token classifications for entity preservation and Sentiment Loss: KL divergence between predicted vs GT DistilBERT sentiment logits	NER Loss: KL divergence between predicted vs GT BERT-NER token classifications for entity preservation. Scene Graph Loss: Symmetric difference ratio of entities and relations between predicted vs GT scene graphs
Metrics	Frechet Inception Distance (FID), Frechet DinoV2 Distance (FDD), CLIP maximum mean discrepancy distance (CMMD), Learned Perceptual Image Patch Similarity (LPIPS), Colorfulness (CF), Peak Signal to Noise Ratio (PSNR)	FBD-4(fbd of average of last 4 hidden layers), Frechet Bert Distance(FBD) (Xiang et al. 2021), BERTScore (Zhang et al. 2020), BLEU-4(Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), COMET (Rei et al. 2020)	FBD-4(fbd of average of last 4 hidden layers), FBD (Xiang et al. 2021), BERTScore (Zhang et al. 2020), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), BLEU-4(Papineni et al. 2002), METEOR (Banerjee and Lavie 2005)
Data Sets	ImageNet (Deng et al. 2009), COCO (Lin et al. 2014), and ADE20K (Zhou et al. 2019) (Zhou et al. 2017)	IWSLT 2014 (Cettolo, Girardi, and Federico 2012) and WMT-14 (Luong and Manning 2015)	COCO (Lin et al. 2014) and flickr30k (Plummer et al. 2015)

Table 2: Experimental configurations for the three selected tasks.

while the training loss remains stable. This could happen for a variety of reasons. For example, consider the FID metric that has gained popularity as an evaluation metric for colorization models. The loss fits a multivariate Gaussian distribution over a feature space extracted from the Inception network. As this feature space is much lower dimensional than the input images, some information is inherently lost. For example, it could be that two colorized images with the same MSE to the ground truth, but different invariant losses, map to the same Inception features. If this occurs, the differences being captured by the invariant loss have been forgotten in the feature space used for evaluation. In this instance, we would say that the Inception features are not *sensitive* to the selected invariant.

As our primary supposition in this work is that invariants are properties that should be approximately preserved by "good" labels, it is critical that the evaluation metrics be sensitive to the selected invariant(s). Our first set of experiments

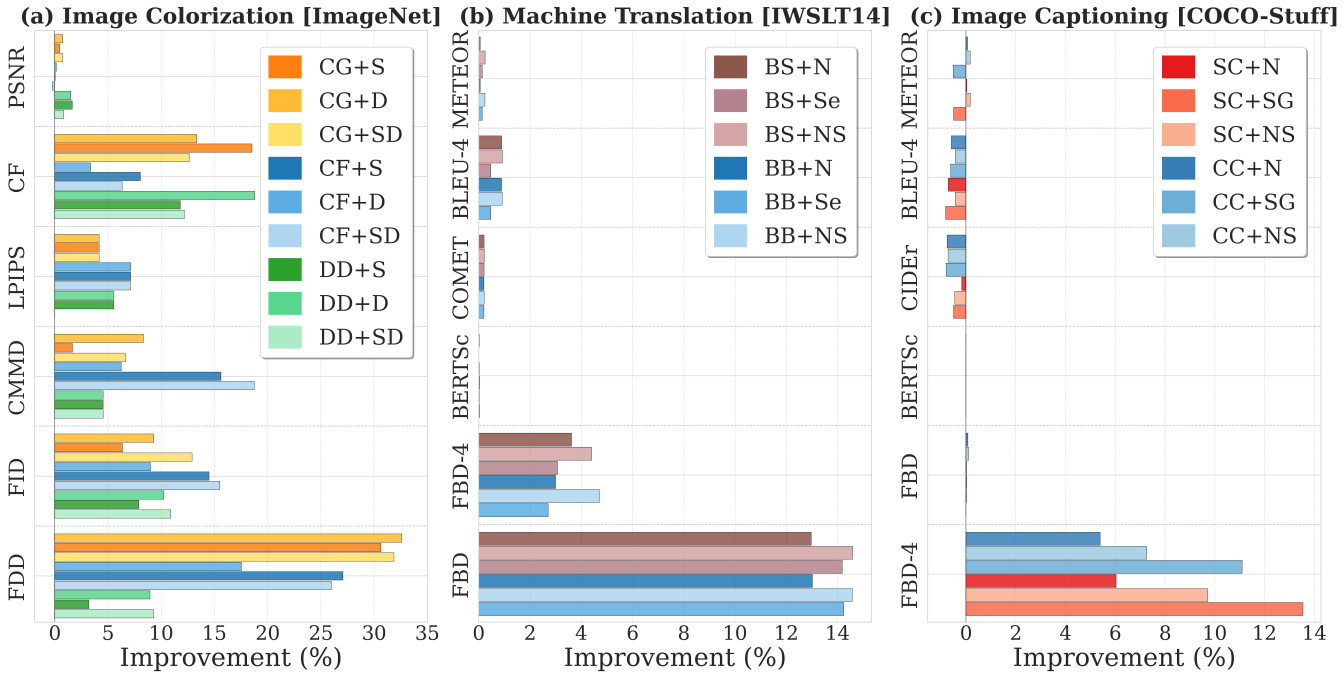


Figure 2: Performance improvement of architectures trained with invariants versus training without invariants, evaluated across common metrics for three tasks. (a) Image Colorization: Base models are ChromaGAN (CG), ColorFormer (CF), and DDColor (DD), each augmented with Segmentation (+S), Depth (+D), or both (+SD). (b) Machine Translation: Base models are Bi-SimCut (BS) and BiBERT (BB), augmented with NER (+N), SeAM (+Se), or both (+NS). (c) Image Captioning: Base models are SmallCap (SC) and ClipCap (CC), augmented with NER (+N), SceneGraph (+SG), or both (+NS). Distributional metrics (FDD, FBD, FID, CMMD) show larger improvements than traditional metrics (PSNR, BLEU-4, METEOR), indicating higher sensitivity to invariant-based training.

is designed to assess the sensitivity of a variety of evaluation metrics to a selected set of invariants for each of our three chosen tasks. To this end, we trained a variety of architectures both with and without an invariant loss and evaluated them using a variety of metrics on validation data.

Figure 2 illustrates the best possible performance improvement (compared against the baseline with no invariant loss during training) for each invariant/metric pair obtained via hyperparameter search on the indicated validation data sets for our chosen tasks. Note that only α, the hyperparameter for the invariant loss was tuned here, all other settings for the architectures remained untouched from the baseline. For the image captioning task, even a small invariant penalty during training results in $\pm 1\%$ performance improvement / degradation for METEOR, BLEU-4, and CIDEr. FBD and BertScore show essentially no differences with respect to training with/without an invariant penalty. This suggests that these metrics are not very sensitive to the selected invariants for this task. However, FBD-4, constructed from the last 4 layers of the BERT network, shows $5\% - 13\%$ improvement across models/invariants. For machine translation, METEOR, BLEU-4, CIDEr, and BERTScore all show a $< 1\%$ improvement while FBD-4 improves $2\% - 5\%$ and FBD improves $13\% - 15\%$. A similar trend can be observed for the colorization task: PSNR, a traditional metric, shows a $< 2\%$ improvement, LPIPS shows moderate $4\% - 7\%$ improvement, FID shows $7\% - 16\%$ improvement, and FDD shows a more dramatic $3\% - 33\%$ improvement. In all cases, metrics derived from richer feature embeddings tend to show increased sensitivity to training with invariants.

From these observations, we can conclude that, while training with an invariant penalty does clearly result in different predictions, many of the standard error metrics for evaluating task performance are almost insensitive to these changes compared with distributional metrics and metrics derived from richer feature spaces. Put another way, if we had valdiated a model using the insensitive metrics, we essentially would not be able to tell whether or not the selected model respects the invariant. This suggests that the natural invariants considered here are not being automatically enforced through the standard train/validate pipeline.

The Effect of Invariants on Task Performance

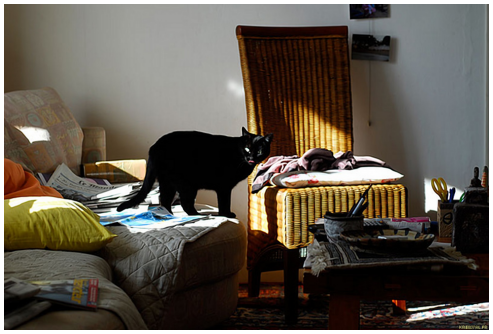
Above, we observed that distributional metrics seem to be more sensitive to training with invariants than their traditional counterparts when we consider the best performance that can be obtained on validation data. This sensitivity is important if we care about preserving the invariants in practice, but high sensitivity alone does not guarantee that a metric is necessarily "good" for evaluating task performance. In particular, models that perform the best along one metric could perform poorly on all other metrics. In this case, high

Method	#P	ImageNet (val50k)						COCO-Stuff(val5k)					
		FDD↓	FID↓	CMMD↓	LPIPS↓	CF↑	PSNR↑	FDD↓	FID↓	CMMD↓	LPIPS↓	CF↑	PSNR↑
ChromaGAN	173.9M	**42.36**	5.81	0.597	0.236	27.41	23.65	94.90	16.73	0.715	0.233	27.89	23.66
+Seg	173.9M	**29.38**	5.44	0.679	0.236	29.36	23.76	81.07	15.50	0.603	0.233	29.56	23.82
+Depth	173.9M	**28.55**	5.27	0.565	0.231	28.73	23.83	79.28	15.29	0.659	0.229	28.95	23.91
+Seg+Depth	173.9M	**28.86**	5.06	0.599	0.231	28.69	23.83	80.00	14.58	0.597	0.229	29.99	23.82
ColorFormer	44.8M	**21.16**	2.00	0.318	0.137	27.57	24.91	59.38	9.03	0.377	0.134	28.30	24.70
+Seg	44.8M	**15.43**	1.71	0.273	0.130	28.29	24.92	52.52	8.49	0.370	0.134	29.21	24.67
+Depth	44.8M	**17.45**	1.82	0.302	0.129	28.16	24.95	55.31	8.80	0.385	0.133	28.83	24.73
+Seg+Depth	44.8M	**15.66**	1.69	0.263	0.131	28.52	24.85	52.74	8.45	0.376	0.135	29.39	24.61
DDColor	55M	**29.60**	4.68	0.436	0.179	52.84	21.22	71.58	12.08	0.484	0.177	50.88	21.43
+Seg	55M	**28.65**	4.31	0.423	0.174	51.14	21.57	69.70	11.48	0.465	0.172	50.10	21.73
+Depth	55M	**26.95**	4.20	0.418	0.174	51.63	21.54	69.69	11.22	0.462	0.172	50.27	21.73
+Seg+Depth	55M	**26.85**	4.17	0.419	0.175	52.12	21.40	69.35	11.20	0.465	0.172	50.67	21.63

Table 3: Quantitative comparison of image colorization methods. ↑ (↓) indicates higher (lower) values are better; Bold values in FDD column of ImageNet indicate the model selection criterion. The Seg architecture is SAN, and depth architecture is Midas.

Model	#P	COCO-Stuff (val5k)						Flickr30k (val1k)					
		FBD-4↓	FBD↓	BERTScore↑	CIDEr↑	BLEU-4↑	METEOR↑	FBD-4↓	FBD↓	BERTScore↑	CIDEr↑	BLEU-4↑	METEOR↑
SmallCap	7M	**62.60**	41.90	0.9224	118.49	36.68	27.90	58.45	39.12	0.9189	112.30	34.82	26.45
+NER	7M	**58.82**	41.89	0.9223	116.29	36.07	27.91	55.20	39.08	0.9187	110.15	34.25	26.48
+SceneGraph	7M	**54.12**	41.90	0.9218	115.47	35.68	27.76	51.85	39.15	0.9182	109.42	33.89	26.32
+NER+SceneGraph	7M	**56.52**	41.89	0.9226	117.48	36.52	27.95	53.78	39.11	0.9191	111.68	34.71	26.51
ClipCap	43M	**48.20**	38.45	0.9301	125.80	39.15	29.20	44.60	36.25	0.9267	119.45	37.28	27.85
+NER	43M	**45.60**	38.42	0.9299	123.45	38.52	29.22	42.15	36.21	0.9265	117.30	36.65	27.88
+SceneGraph	43M	**42.85**	38.44	0.9295	122.30	38.01	29.05	39.85	36.28	0.9260	116.52	36.14	27.72
+NER+SceneGraph	43M	**44.70**	38.41	0.9303	124.90	38.98	29.25	41.25	36.24	0.9269	118.85	37.12	27.91

Table 4: Performance on image captioning benchmarks. Bold values in FBD-4 column of COCO-Stuff indicate the model selection criterion.



Ground Truth

Caption: A cat standing on a couch in a cluttered living room.

Entities:

Head	Span	Modifiers
cat	a cat	a
couch	a couch	a
living room	a cluttered living room	a, cluttered

Relations:

Subject	Relation	Object
cat	on	couch
cat	in	living room

SmallCap+NER+SceneGraph

Caption: A black cat standing on a couch in a living room.

Entities:

Head	Span	Modifiers
cat	a black cat	a, black
couch	a couch	a
living room	a living room	a

Relations:

Subject	Relation	Object
cat	on	couch
cat	in	living room

SmallCap

Caption: A black cat standing on top of a bed.

Entities:

Head	Span	Modifiers
cat	a black cat	a, black
bed	a bed	a

Relations:

Subject	Relation	Object
cat	on	bed

Figure 3: Image captions produced by models trained with/without scenegraph and NER invariants for an image of a cat (left).

performance was achieved along one metric at the expense of the other metrics. In this section, we explore whether or not training with an invariant penalty yields performance improvements more broadly. That is, does incorporating task invariants as above and validating with an invariant-sensitive metric yield practical performance gains?

For each of our tasks, we trained a variety of architectures with different invariant losses and report the resulting performance across common task metrics. The best model for each architecture/invariant combination was selected using the most sensitive metric for each task from the previous section, i.e., FDD for colorization, FBD-4 for image captioning, and FBD for machine translation, on a validation data set. The results of our investigation can be found in Tables 3 and 4. As was noted above, the best performing models each perform better with respect to the metric used for model selection. However, the more detailed results in the tables illustrate that the performance improvements are not

Model	#P	IWSLT-14						WMT-14					
		FBD↓	FBD-4↓	BERTScore↑	BLEU-4↑	METEOR↑	COMET↑	FBD↓	FBD-4↓	BERTScore↑	BLEU-4↑	METEOR↑	COMET↑
Bi-SimCut	65M	**9.87**	102.15	0.8871	37.83	0.6412	0.5743	15.24	135.42	0.8952	34.47	0.5836	0.6124
+NER	65M	**8.42**	101.89	0.8867	36.95	0.6358	0.5739	14.18	134.76	0.8948	33.82	0.5794	0.6119
+SeAM	65M	**8.51**	103.14	0.8868	37.02	0.6361	0.5737	14.26	136.23	0.8949	33.89	0.5798	0.6117
+NER+SeAM	65M	**8.38**	102.67	0.8869	37.18	0.6373	0.5741	14.09	135.58	0.8950	34.06	0.5809	0.6121
BiBERT	110M	**8.99**	100.24	0.8930	38.20	0.6597	0.5899	13.76	133.18	0.9012	35.24	0.6023	0.6287
+NER	110M	**7.71**	100.03	0.8925	37.35	0.6538	0.5894	12.84	132.91	0.9008	34.52	0.5976	0.6282
+SeAM	110M	**7.82**	101.27	0.8926	37.41	0.6536	0.5890	12.93	134.17	0.9009	34.58	0.5974	0.6279
+NER+SeAM	110M	**7.68**	100.80	0.8927	37.58	0.6548	0.5897	12.76	133.75	0.9010	34.74	0.5987	0.6284

Table 5: Performance on machine translation. Bold values indicate the model selection criterion.

accompanied by significant degredation in the other metrics. For example, in Table 3, Colorformer+Seg achieves the best FDD score across all architectures and yields performance improvements on nearly all metrics across the two validation data sets. Contrast this with the results in Table 4 in which dramatic improvements in FBD-4 are met with minor $\pm 2\%$ differences in nearly all other metrics. A similar observation can be made for FBD in machine translation.

The tables also illustrate the effect of using more than one invariant penalty during training. Recall that, as the invariant networks are frozen during training, the only additional training overhead comes from evaluating these networks. In colorization, ColorFormer+Seg+Depth acheives a lower FDD than ColorFormer+Depth, but is slightly worse than ColorFormer+Seg. A similar observation can be made for ClipClap+NER+Scenegraph for image captioning. In machine translation, BiBERT+NER+SeAM achieves the best overall FBD score. Overall, while adding two potentially conflicting losses can lead to optimization challenges, the results still remain competitive while encouraging preservation of more than one invariant.

Qualitatively, we have found that training with invariants can help correct errors that violate the invariants. In Figure 1, we provide an example of colorization outputs when trained with/without a depth or a segmentation invariant. In both cases we observe that better aligned outputs from the invariant networks correlate with better colorizations. In Figure 3, an example of image captioning with an NER+SceneGraph invariant shows that a better caption is produced and the SceneGraph generated from the ground truth image better matches the SceneGraph from the NER+SceneGraph invariant model (the invaraint-trained model correctly identifies the couch).

Conclusion

In summary, our experimental results illustrate that (1) traditional metrics, e.g., PSNR, BLEU-4, etc., show minimal sensitivity to invariants, (2) distributional metrics, e.g., FID, FDD, and FBD variants, show substantial improvements when invariants are enforced without significant loss with respect to traditional metrics, and (3) richer feature representations, e.g., DinoV2 in FDD, show even greater sensitivity to invariants than less rich representations. Critically, our metric evaluation framework does not depend on human evaluators, which makes this a flexible approach that extends beyond the tasks considered here. Furthermore, our results suggest that future metric design should favor richer feature encoders, e.g., DinoV2 rather than Inception for images, and consider incorporating explicit invariant-violation penalties.

Limitations and Future Work

Our framework relies on approximate task invariants derived from pretrained models, which may introduce biases or fail in low-resource domains where such models are unavailable. Moreover, we do not directly validate the hypothesis that enforcing approximate invariants leads to improved human evaluation; instead, we position invariants as a scalable diagnostic when human studies are impractical. Finally, our empirical study focuses on three domains (image colorization, machine translation, and image captioning). Extending invariants-as-probes to other multi-output structured tasks and integrating them directly into metric design are interesting directions for future work.

Acknowledgements

This work was supported in part by NSF grant 2327245.

References

Ahmed, K.; Teso, S.; Chang, K.-W.; Van den Broeck, G.; and Vergari, A. 2022. Semantic Probabilistic Layers for Neuro-symbolic Learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*. Red Hook, NY, USA: Curran Associates Inc.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, 382–398. Berlin, Heidelberg: Springer-Verlag.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72. Kerrville, TX, USA: Association for Computational Linguistics.

Cettolo, M.; Girardi, C.; and Federico, M. 2012. Wit3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the Conference of European Association for Machine Translation (EAMT)*, 261–268. Switzerland: European Association for Machine Translation.

Chong, M. J.; and Forsyth, D. 2020. Effectively Unbiased FID and Inception Score and Where to Find Them. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6069–6078. Los Alamitos, CA, USA: IEEE Computer Society.

Cohen, T.; and Welling, M. 2016. Group Equivariant Convolutional Networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML)*, 2990–2999. JMLR.org.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A Large-scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255. Los Alamitos, CA, USA: IEEE Computer Society.

Gao, P.; He, Z.; Wu, H.; and Wang, H. 2022. Bi-SimCut: A Simple Strategy for Boosting Neural Machine Translation. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies*, 3938–3948. Kerrville, TX, USA: Association for Computational Linguistics.

Hasler, D.; and Suesstrunk, S. E. 2003. Measuring Colorfulness in Natural Images. In *Human Vision and Electronic Imaging (HVEI) VIII*, volume 5007, 87–95. International Society for Optics and Photonics, Bellingham, WA, USA: SPIE.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-scale Update Rule Converge to a Local Nash Equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 6629–6640. Red Hook, NY, USA: Curran Associates Inc.

Jayasumana, S.; Ramalingam, S.; Veit, A.; Glasner, D.; Chakrabarti, A.; and Kumar, S. 2024. Rethinking FID: Towards a Better Evaluation Metric for Image Generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9307–9315. Los Alamitos, CA, USA: IEEE Computer Society.

Ji, X.; Jiang, B.; Luo, D.; Tao, G.; Chu, W.; Xie, Z.; Wang, C.; and Tai, Y. 2022. ColorFormer: Image Colorization via Color Memory Assisted Hybrid-attention Transformer. In *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, 20–36. Berlin, Heidelberg: Springer-Verlag.

Kang, X.; Yang, T.; Ouyang, W.; Ren, P.; Li, L.; and Xie, X. 2023. DDcolor: Towards photo-realistic image colorization via dual decoders. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 328–338. Los Alamitos, CA, USA: IEEE Computer Society.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the 13th European Conference on Computer Vision (ECCV)*, 740–755. Berlin, Heidelberg: Springer-Verlag.

Luong, M.-T.; and Manning, C. D. 2015. Stanford Neural Machine Translation Systems for Spoken Language Domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, 76–79. Kerrville,TX, USA: Association for Computational Linguistics.

Misra, I.; Shrivastava, A.; Gupta, A.; and Hebert, M. 2016. Cross-Stitch Networks for Multi-Task Learning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (ICCV)*, 3994–4003. Los Alamitos, CA, USA: IEEE Computer Society.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip Prefix for Image Captioning. arXiv:2111.09734.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318. Kerrville, TX, USA: Association for Computational Linguistics.

Parmar, G.; Zhang, R.; and Zhu, J.-Y. 2022. On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *2015 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2641–2649. Los Alamitos, CA, USA: IEEE Computer Society.

Ramos, R.; Martins, B.; Elliott, D.; and Kementchedjhieva, Y. 2023. Smallcap: Lightweight Image Captioning Prompted with Retrieval Augmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2840–2849. Los Alamitos, CA, USA: IEEE Computer Society.

Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2020. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Rei, R.; Stewart, C.; Farinha, A. C.; and Lavie, A. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. Kerrville, TX, USA: Association for Computational Linguistics.

Ruder, S. 2017. An Overview of Multi-task Learning in Deep Neural Networks. arXiv:1706.05098.

Schuster, S.; Krishna, R.; Chang, A.; Fei-Fei, L.; and Manning, C. D. 2015. Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval. In *Proceedings of the Fourth Workshop on Vision and Language*, 70–80. Kerrville, TX, USA: Association for Computational Linguistics.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. CIDEr: Consensus-based Image Description Evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4566–4575. Los Alamitos, CA, USA: IEEE Computer Society.

Vitoria, P.; Raad, L.; and Ballester, C. 2020. ChromaGAN: Adversarial Picture Colorization with Semantic Class Distribution. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2445–2454. Los Alamitos, CA, USA: IEEE Computer Society.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's Transformers: State-of-the-Art Natural Language Processing. arXiv:1910.03771.

Worrall, D. E.; Garbin, S. J.; Turmukhambetov, D.; and Brostow, G. J. 2017. Harmonic Networks: Deep Translation and Rotation Equivariance. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5028–5037. Los Alamitos, CA, USA: IEEE Computer Society.

Wu, H.; Mao, J.; Zhang, Y.; Jiang, Y.; Li, L.; Sun, W.; and Ma, W.-Y. 2019. Unified Visual-semantic Embeddings: Bridging Vision and Language with Structured Meaning Representations. In *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6609–6618. Los Alamitos, CA, USA: IEEE Computer Society.

Xiang, J.; Liu, Y.; Cai, D.; Li, H.; Lian, D.; and Liu, L. 2021. Assessing Dialogue Systems with Distribution Distances. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2192–2198. Kerrville, TX, USA: Association for Computational Linguistics.

Xu, H.; Van Durme, B.; and Murray, K. 2021. BERT, mBERT, or BiBERT? A Study on Contextualized Embeddings for Neural Machine Translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Kerrville, TX, USA: Association for Computational Linguistics.

Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; and Bai, X. 2023. Side Adapter Network for Open-vocabulary Semantic Segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence (TPAMI)*, 15546–15561.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 586–595. Los Alamitos, CA, USA: IEEE Computer Society.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*. Appleton, WI, USA: International Conference on Learning Representations.

Zhao, S.; Ma, H.; and Ermon, S. 2018. Bias and Generalization in Deep Generative Models: An Empirical Study. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 10815–10824. Red Hook, NY, USA: Curran Associates Inc.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene Parsing through ADE20K Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 633–641. Los Alamitos, CA, USA: IEEE Computer Society.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic Understanding of Scenes Through the ADE20k Dataset. *International Journal of Computer Vision*, 127(3): 302–321.

Zhu, J.-Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A. A.; Wang, O.; and Shechtman, E. 2017. Toward Multimodal Image-to-image Translation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 465–476. Red Hook, NY, USA: Curran Associates Inc.