SceneTrilogy: On Scene Sketches and its Relationship with Text and Photo

Pinaki Nath Chowdhury[1,2] Ayan Kumar Bhunia[1]
Tao Xiang[1,2] Yi-Zhe Song[1,2]
[1]SketchX, CVSSP, University of Surrey, United Kingdom.
[2]iFlyTek-Surrey Joint Research Centre on Artificial Intelligence.
{p.chowdhury, a.bhunia, t.xiang, y.song}@surrey.ac.uk

Abstract. We for the first time extend multi-modal scene understanding to include that of free-hand scene sketches. This uniquely results in a trilogy of scene data modalities (sketch, text, and photo), where each offers unique perspectives for scene understanding, and together enable a series of novel scene-specific applications across discriminative (retrieval) and generative (captioning) tasks. Our key objective is to learn a common three-way embedding space that enables *many-to-many* modality interactions (e.g, sketch+text $\rightarrow$ photo retrieval). We importantly leverage the information bottleneck theory to achieve this goal, where we (i) decouple *intra-modality* information by minimising the mutual information between modality-specific and modality-agnostic components via a conditional invertible neural network, and (ii) align *cross-modalities information* by maximising the mutual information between their modality-agnostic components using InfoNCE, with a specific multihead attention mechanism to allow many-to-many modality interactions. We spell out a few insights on the complementarity of each modality for scene understanding, and study for the first time a series of scene-specific applications like joint sketch- and text-based image retrieval, sketch captioning.

1 Introduction

Scene understanding sits at the very core of computer vision. As research matures on object-level understanding [18, 23], an encouraging shift has been witnessed in recent years on the understanding of scenes, e.g., scene recognition [93], scene captioning [48], scene synthesis [24], and scene retrieval [11, 49].

The development of scene research has largely progressed from that of single modality [93, 94] to the very recent focus on multi-modality [11, 5]. The latter setting not only triggered a series of practical applications [87, 24, 49, 95] but importantly helped to cast insights into scene understanding on a conceptual level (i.e., what is really being perceived by humans). To date, research on multi-modal scene understanding has mainly focused on two modalities – text and photo [51,54,53]. Through applications such as text-based scene retrieval (TBIR) [26] and scene captioning [54, 53, 17], a deeper research question was asked: To what level can text represent the human interpretation of a scene?

In this paper, we follow this trend of multi-modal scene understanding and importantly set out to further our understanding of different forms of scene

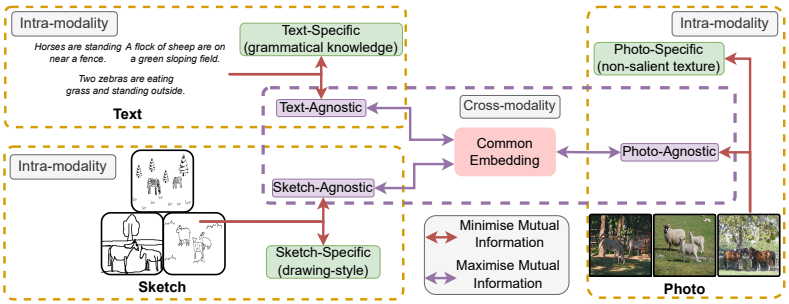

Fig. 1. We learn a common three-way embedding space that enables *many-to-many* modality interactions. To decouple the intra-modality information we minimise mutual information between (text-specific, sketch-specific, photo-specific) and (text-agnostic, sketch-agnostic, photo-agnostic) components. We align cross-modalities by maximising mutual information between the agnostic components (text-agnostic, sketch-agnostic, photo-agnostic) using multihead attention on top of InfoNCE.

representation (i.e., other than just text and photo). Our main contribution is the introduction of an entirely new modality for scene understanding, that of scene sketches in context of a trilogy of scene representations – sketch, text, and photo). However, achieving this trilogy is non-trivial due to the distinct differences across sketch, text, and photo modalities. Naively mapping the information extracted from all three modalities does not fully accommodate for their complementarity [50]. This is because sketch, text, and photo modalities are not necessarily isomorphic i.e., while sketch inherently encode object pose and structure, it leaves other details open, like colour which comes naturally for text and photo. It is well-acknowledged that existing works specifically aligning text and photo (i.e., two modalities), such as CAMP [86] or CLIP [61], perform well on tasks that dictate a common embedding (e.g., retrieval) but fails to generalise to ones that require modality-specific information (e.g., captioning) [53]. On the other hand, more general cross modality works with three or more modalities [5, 75] typically use a naive triplet-based loss without modeling the interactions across modalities to support tasks such as joint sketch- and text-based image retrieval.

To fully explore the unique attributes of each modality and to underpin a many-to-many modality interactions, we desire an embedding that (i) knows modality-specific information, e.g., grammatical knowledge for text, and different drawing styles for sketch, and texture information of non-salient photo regions (ii) has the flexibility in exploring the complementarity of modality-agnostic information extracted from individual modalities in a many-to-many manner, and in turn support novel tasks such as joint sketch- and text-based image retrieval (relating two modalities, i.e., sketch and text with one, i.e., photo).

In this paper, we overcome the aforementioned challenges by leveraging an interpretation of information bottleneck to learn a shared embedding space across the trilogy of sketch, text, and photo. Fig. 1 offers a schematic of the overall learning framework. We start by conducting *intra-modality* reasoning to decouple modality-agnostic information and modality-specific information, by minimising

their mutual information using a conditional invertible neural network [4, 64]. We then combine the modality-agnostic information from each modality to reason around the complementary information residing in each modality. This is achieved by maximising the mutual information across sketch-agnostic, photo-agnostic and text-agnostic feature using InfoNCE [56, 74]. We importantly leverage a multihead attention mechanism [40] on top of InfoNCE to model many-to-many associations across the three modalities.

In summary, our contributions are: (i) We extend multi-modal scene understanding to include that of free-hand scene sketches. (ii) This uniquely results in a trilogy of scene modalities (sketch, text, and photo) by learning a common three-way embedding space that enables many-to-many modality interactions. (iii) We leverage information bottleneck to achieve this goal, where we decouple intra-modality information between modality-specific and modality-agnostic components using conditional invertible neural networks, (iv) to align cross-modalities information for many-to-many modality interactions we use InfoNCE with a multihead attention mechanism.

2 Related Work

2.1 Sketch Understanding

Early efforts were focused on sketch recognition [23, 30, 70, 42]. It started with hand-crafted representations such as bag-of-words [23] and Fisher Vector [70] to deep neural networks with impressive outcomes like sketch recognition methods outperforming humans [60]. The research focus, however, has been predominantly on single-object sketches, via applications such as sketch recognition [23], sketch segmentation [78, 32, 71, 69], sketch synthesis [25, 9], or in a cross-modal setting such as fine-grained sketch-based photo retrieval [10, 67], sketch-based photo editing [58], sketch-based image generation [12]. As research matures, recent works naturally took a step towards *scene*-level for deeper and richer reasoning about sketched visual forms [95, 24, 49]. Zou *et al.* [95] studied segmentation and colourisation on scene-level sketches. Gao *et al.* [24] proposed scene-sketch to image generation via a generative adversarial approach in two sequential modules. Liu *et al.* [49] investigated fine-grained scene sketch-based image retrieval using graph convolutional networks. However, existing works are primarily limited to understanding scene sketches and its relation with photos. Prior work [75] has shown the complimentarity of sketches, text, and photos on object-level. We explore this complementarity for the more complex and realistic setup of scene understanding with the introduction of *SceneTrilogy*.

2.2 Cross-Modal Mapping

Most contemporary cross-modal work focused on text and photo, via image captioning [85, 87, 3]. In particular splitting information into modality-specific and shared agnostic representation using conditional invertible neural network has shown to benefit cross-modal transfer [64], where Mahajan *et al.* [53] employed

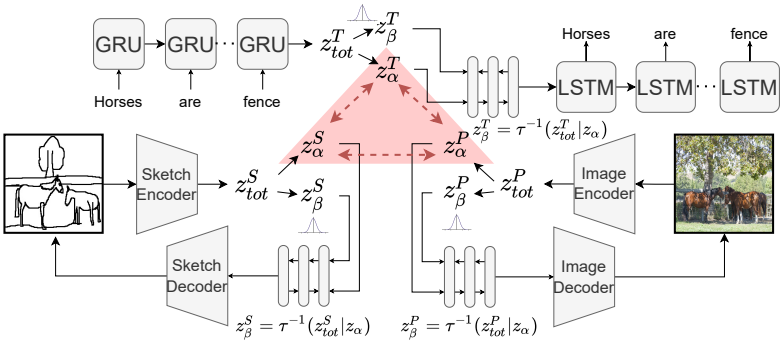

Fig. 2. Schematic representation of the proposed model for *SceneTrilogy*. Our unified model overcomes the extreme modality gap between Sketch, Text, and Photo by splitting information into modality-agnostic and modality-specific component using conditional Invertible Neural Network. The modality-agnostic components are combined using multihead attention mechanism. Finally, to align the modality-agnostic components from sketch, text, and photo, we employ InfoNCE.

normalising flow [20] to bridge the modality gap. A popular approach for maximising cross-modal mutual information is InfoNCE [56], albeit specifically for two modality problems [31, 91, 79]. Recently, the multihead attention mechanism [40] has found a new application in combining information from multiple modalities. Perceiver [33] shows remarkable ability to understand information from image, text, and point cloud for the task of classification. In this paper, to bridge the extreme modality gap between sketch, text, and image, we split the information encoded from each modality using conditional invertible neural network [4, 64], followed by combining different query modalities like sketch and text using multihead attention [40]. Finally, we align query modalities with target (e.g., image) modalities using InfoNCE [56] by maximising mutual information. Further discussion on relevant work is included in the supplementary.

3 Methodology

We propose a unified model (illustrated in Fig. 2) that can model the joint distribution of Sketch, Text, and Photo – SceneTrilogy. The resulting model can perform several downstream tasks (i) fine-grained sketch-based image retrieval (FG-SBIR), (ii) fine-grained text-based image retrieval (FG-TBIR), (iii) fine-grained sketch+text based image retrieval (FG-STBIR), (v) image captioning, (vi) sketch captioning. Mapping feature representations across modalities using mean-squared distance is sub-optimal as (i) Such naive approaches do not decouple modality-specific information from modality-agnostic component. (ii) In addition, unimodal losses such as mean squared error cannot preserve high-dimensional scene information across sketch, text, and photo [56]. Hence, in this paper, we use information bottleneck principles to decouple modality-agnostic from modality-specific information. The modality-agnostic components from sketch, text, and photo are aligned together using a combination of multihead attention [40] and InfoNCE [56]. First, we provide a background of the

retrieval framework followed by a brief overview of captioning. Next, we introduce the training setup of SceneTrilogy. Finally, we show how the trained unified model can be used to perform FG-STBIR and sketch captioning.

3.1 Background

Background Retrieval Framework: First, we briefly summarize a baseline retrieval framework that remains state-of-the-art for fine-grained sketch-based image retrieval (FGSBIR) literature to date. Given a query-target pair (e.g., sketch-image) represented as (Qr, Tr), a feature extractor $\mathcal{F}_\theta(\cdot)$ parameterised by θ is used to get the feature map $\mathrm{U} = \mathcal{F}_\theta(\mathrm{Qr}) \in \mathbb{R}^{h_{Qr} \times w_{Qr} \times c}$ and $\mathrm{V} = \mathcal{F}_\theta(\mathrm{Tr}) \in \mathbb{R}^{h_{Tr} \times w_{Tr} \times c}$. The encoder $\mathcal{F}_\theta$ can be modelled by CNN [88], LSTM [42], Transformer [46], Graphs [49,59,77], or their combination [8]. We perform Global Average Pooling (GAP) [47] on the backbone output feature-map to obtain final feature representation $\mathrm{U} : \mathbb{R}^{h_{Qr} \times w_{Qr} \times c} \rightarrow \mathbb{R}^c$ and $\mathrm{V} : \mathbb{R}^{h_{Tr} \times w_{Tr} \times c} \rightarrow \mathbb{R}^c$. For training, the distance to query anchor (a) from negative target (n), denoted as $\beta^- = ||\mathcal{F}(a) - \mathcal{F}(n)||_2$ should increase while that from the positive target (p), $\beta^+ = ||\mathcal{F}(a) - \mathcal{F}(p)||_2$ should decrease. Training is achieved via triplet loss with a margin $\mu > 0$ as a hyperparameter:

$$\mathrm{L}^{trip} = max\{0, \mu + \beta^+ - \beta^-\} \tag{1}$$

Background Captioning Framework: Next, we briefly introduce a popular captioning framework [87,51]. Given a query Qr (e.g., sketch or image), the goal of captioning is to generate a textual description $\mathrm{T} = \{w_1, w_2, \ldots, w_T\}$, where the ground-truth is denoted by $\mathrm{T}^* = \{w_1^*, w_2^*, \ldots, w_T^*\}$. A feature extractor $\mathcal{F}_\theta$ parameterised by θ is used to encode the query Qr to get the feature map $\mathrm{U} = \mathcal{F}_\theta(Qr) \in \mathbb{R}^{h_{Qr} \times w_{Qr} \times c}$, followed by GAP to get $\mathrm{U} : \mathbb{R}^{h_{Qr} \times w_{Qr} \times c} \rightarrow \mathbb{R}^c$. The decoder $p_\varphi(\cdot)$, parameterised by φ, predicts the textual description by generating one word at every time step conditioned on U and all previously generated words. For training, we maximise the likelihood as:

$$\mathrm{L}^{CE} = -\sum_{t=1}^{T} log(p_\varphi(w_t = w_t^* | \mathrm{U}, w_1^*, \ldots, w_{t-1}^*) \tag{2}$$

3.2 Training for SceneTrilogy

SceneTrilogy models the joint distribution of Sketch, Text, and Photo. To estimate this joint distribution, we extend the (variational) lower bound on the marginal likelihood defined in [38] as $\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}[-\log q_\phi(z|x) + \log p_\theta(x, z)]$ to the setup of three modalities as:

$$\begin{aligned} \log p_\theta(x_S, x_T, x_P) \geq & \mathbb{E}_{q_\phi(z|x_S, x_T, z_P)}[\log p_\theta(x_S, x_T, x_P|z)] \\ & + \mathbb{E}_{q_\phi(z|x_S, x_T, z_P)}[\log p_\theta(z) - \log q_\phi(z|x_S, x_T, x_P)] \end{aligned} \tag{3}$$

where $p_\theta(x_S), p_\theta(x_T), p_\theta(x_P)$ denotes the data distribution of sketch, text, and image, parameterised by θ. The variational posterior $q_\phi(z|x_S, x_T, x_P)$ is parameterised by ϕ, where z is the latent variable. The first expectation term $p_\theta(x_S, x_T, x_P|z)$ is known as the reconstruction error and is solved using estimation by sampling [38], represented as L_{rec}. In the following sections, we solve the

second expectation term comprising of $p_\theta(z)$ and $q_\phi(z|x_S, x_T, x_P)$ by factorising the latent variable z into sketch, text, and photo components.

Factorising the latent variable: Our objective is to factorise the latent space into modality-agnostic and modality-specific components. The latent variable z comprise of $z^S_{tot}, z^T_{tot}, z^P_{tot}$ representing the latent variable for sketch, text, and photo modalities. Each of $\{z^S_{tot}, z^T_{tot}, z^P_{tot}\} \in \mathbb{R}^c$ comprise of a modality-agnostic and modality-specific component represented as: $z^S_{tot} = [z^S_\alpha, z^S_\beta]$; $z^T_{tot} = [z^T_\alpha, z^T_\beta]$; and $z^P_{tot} = [z^P_\alpha, z^P_\beta]$ where $\{z^S_\alpha, z^T_\alpha, z^P_\alpha\} \in \mathbb{R}^d$ are the modality-agnostic components and $\{z^S_\beta, z^T_\beta, z^P_\beta\} \in \mathbb{R}^{c-d}$ are the modality-specific components. (i) Since the modality-agnostic is transferable across modalities, we assume them to represent a common/shared information as $z^S_\alpha \simeq z^T_\alpha \simeq z^P_\alpha \simeq z_\alpha$. (ii) Similarly, the modality-specific information is not shared across modalities and captures unique properties of a particular modality (e.g., grammatical knowledge in text modality). Hence, we can assume the modality-specific information to be conditionally independent, as: $p(z^S_\beta, z^T_\beta, z^P_\beta) \simeq p(z^S_\beta) \cdot p(z^T_\beta) \cdot p(z^P_\beta)$. Therefore, the latent z in Eq. 3 can be factorised as:

$$\begin{aligned}
&\mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log q_\phi(z|x_S, x_T, x_P) - \log p_\theta(z)] = \\
&D_{KL}(q_{\phi_\alpha}(z_\alpha|x_S, x_T, x_P)||p_{\theta_\alpha}(z_\alpha)) + D_{KL}(q_{\phi_S}(z^S_\beta|x_S, z_\alpha)||p_{\theta_S}(z^S_\beta|z_\alpha)) \\
&+ D_{KL}(q_{\phi_T}(z^T_\beta|x_T, z_\alpha)||p_{\theta_T}(z^T_\beta|z_\alpha)) + D_{KL}(q_{\phi_P}(z^P_\beta|x_P, z_\alpha)||p_{\theta_P}(z^P_\beta|z_\alpha))
\end{aligned} \tag{4}$$

Proof: See Supplementary material.

The above equation has four KL-divergence terms. The three terms analogous to $D_{KL}(q_{\phi_S}(z^S_\beta|x_S, z_\alpha)||p_{\theta_S}(z^S_\beta|z_\alpha))$ models the modality-specific information in sketch, text, and photo modalities that is estimated using information bottleneck principles. The fourth term $D_{KL}(q_{\phi_\alpha}(z_\alpha|x_S, x_T, x_P)||p_{\theta_\alpha}(z_\alpha))$ models the modality-agnostic information that is transferable across modalities and computed using a combination of multihead attention [40] and InfoNCE [56].

Interpretation as Information Bottleneck: Information bottleneck aims to extract minimal information from the input variable that can sufficiently represent the output variable i.e., it tries to establish a tradeoff between compression and prediction [80, 1, 81]. We leverage this principle to disentangle or split modality-agnostic information from modality-specific component. Our objective is to learn modality-agnostic $\{z^S_\alpha, z^T_\alpha, z^P_\alpha\}$ that is maximally transferable across modalities while being minimally informative about modality-specific $\{z^S_\beta, z^T_\beta, z^P_\beta\}$. We begin by first estimating the posterior $q_{\phi_T}(z^T_\beta|x_T, z_\alpha)$ in $D_{KL}(q_{\phi_T}(z^T_\beta|x_T, z_\alpha)||p_{\theta_T}(z^T_\beta|z_\alpha))$ using a modality specific encoder $\mathcal{F}_{\phi_S}$. To estimate the conditional prior $p_{\theta_T}(z^T_\beta|z_\alpha)$, we employ a conditional neural network (cINN) $\tau(\cdot)$[1]. Although one can estimate the conditional priors using a VAE-like multivariate Gaussian [68], recent literature [53, 64] suggests that such simplified assumption is sub-optimal. For text, the cINN can reconstruct the total information z^T_{tot} given the modality-agnostic z_α and modality-specific z^T_β as: $z^T_{tot} = \tau(z^T_\beta|z_\alpha)$. During training, we learn the prior distribution of modality-

[1] We learn three cINNs for sketch, text, and photo modalities.

specific information $q(z_\beta^T) = \mathcal{N}(z_\beta^T, 0, \mathbf{I})$ as: $z_\beta^T = \tau^{-1}(z_{tot}^T|z_\alpha)$. We aim to minimise the mutual information between z_β^T and z_α that is usually defined as [1,64]:

$$\min\ \mathcal{I}(z_\beta^T, z_\alpha) \leq \mathbb{E}_{z_\beta^T, z_\alpha} \log \frac{p(z_\beta^T|z_\alpha)}{q(z_\beta^T)} = \mathbb{E}_{z_{tot}^T, z_\alpha} \log \frac{p(\tau^{-1}(z_{tot}^T|z_\alpha)|z_\alpha)}{q(\tau^{-1}(z_{tot}^T|z_\alpha))} \tag{5}$$

Using change-of-variable technique [20], we can express the probability distribution of one variable $p(x)$ using some other variable $p(y)$ under some invertible transformation function $g : X \to Y$ as:

$$p(x) = p(y)|\mathrm{det}J_g(x)| = p(g(x))|\mathrm{det}J_g(x)| \tag{6}$$

where, $\mathrm{det}J_g$ denotes the determinant of the Jacobian of g. Thus we rewrite as,

$$p(\tau^{-1}(z_{tot}^T|z_\alpha)|z_\alpha) = p(z_{tot}^T|z_\alpha)\ |\mathrm{det}J_{\tau^{-1}}(z_{tot}^T|z_\alpha)|^{-1} \tag{7}$$

Substituting Eq. 7 in Eq. 5 we get our required disentanglement loss objective,

$$\begin{aligned} \min_\tau \mathcal{I}(z_\beta^T, z_\alpha) \leq \mathbb{E}_{z_{tot}^T, z_\alpha}\{\log p(\tau^{-1}(z_{tot}^T|z_\alpha)|z_\alpha) - \log q(\tau^{-1}(z_{tot}^T|z_\alpha))\} \\ \mathbb{E}_{z_{tot}^T, z_\alpha}\{-\log q(\tau^{-1}(z_{tot}^T|z_\alpha)) - \log|\mathrm{det}J_{\tau^{-1}}(z_{tot}^T|z_\alpha)| + \log p(z_{tot}^T|z_\alpha)\} \end{aligned} \tag{8}$$

The term $\log p(z_{tot}^T|z_\alpha)$ is constant with respect to τ and hence ignored from the minimisation objective. In summary, we learn a cINN such that it minimises the mutual information between modality-specific z_β^T and modality-agnostic z_α while simultaneously representing the total information $z_{tot}^T = \tau(z_\beta^T|z_\alpha)$ to achive decoupling or *split* of information. Additionally, using invertibility of τ we can also estimate the prior distribution $p_{\theta_T}(z_\beta^T|z_\alpha)$ as $z_\beta^T = \tau^{-1}(z_{tot}^T|z_\alpha)$.

Combination of multiple modalities: SceneTrilogy comprise of three modalities, of which sketch and text can serve as query modalities for retrieval tasks. Unfortunately, current systems [88, 61] allows the user to either sketch or write text for image retrieval. However, while some scene information are easily described via sketch, others such as colour are best represented via text. Moreover, object-level sketch and text has shown to compliment each other to improve performance [75]. But how to design a module that can work on either sketch, or text, or both sketch and text? We solve this by considering the modality-agnostic information from sketch and text $\{z_\alpha^S, z_\alpha^T\}$ be elements of a *set*. Then by extracting a feature representation of this set, we can model either sketch, or text, or both. In particular, feature representation of a set is independent to the number of elements in the set. Hence, our set can contain either one element (sketch or text) or two elements (sketch and text)[2]. To extract feature representation of the set of modality-agnostic components, we employ Lee *et al.* [40] that leverages a multihead attention block (MAB). Let $\eta \in \mathbb{R}^{2\times d}$ represent $\{z_\alpha^S, z_\alpha^T\} \in \mathbb{R}^d$ arranged row-wise in a matrix. We learn a seed vector $\mathcal{V} \in \mathbb{R}^{1\times d}$ such that,

$$\begin{aligned} z^\eta = MAB(\mathcal{V}, rFF(\eta)) = \mathrm{LayerNorm}(H + rFF(H)), \\ \text{where, } H = \mathrm{LayerNorm}(\mathcal{V} + \mathrm{Multihead}(\mathcal{V}, \eta, \eta; \lambda, \omega)) \end{aligned} \tag{9}$$

[2] The set can also contain three elements $(z_\alpha^S, z_\alpha^T, z_\alpha^P)$.

where rFF($\cdot$) denotes row-wise feedforward layer that processes each instance independently and identically, LayerNorm is layer normalisation [6], λ is the learnable parameters of Multihead($\cdot$), and $\omega(\cdot) = softmax(\cdot/\sqrt{d})$ is scaled softmax. Multihead($\cdot$) is an extension of the vanilla attention scheme that computes h attentional heads as,

$$\begin{aligned} \text{Multihead}(\mathcal{V}, \eta, \eta; \lambda, \omega) &= [O_1, O_2, \ldots, O_h]W^O \\ \text{where, } O_j &= Att(\mathcal{V}W_j^Q, \eta W_j^K, \eta W_j^V; \omega);\ \lambda = \{W_j^Q, W_j^K, W_j^V\}_{j=1}^h \end{aligned} \tag{10}$$

$\text{Att}(Q, K, V; \omega) = \omega(QK^T)V$. Hence, using MAB module, we can convert a set representation $\eta \in \mathbb{R}^{1\times d}$ or $\eta \in \mathbb{R}^{2\times d}$ into a feature representation $z^\eta \in \mathbb{R}^{1\times d}$.

Aligning high-dimensional features: In this section, we shall compute the KL-divergence term $D_{KL}(q_{\phi_\alpha}(z_\alpha|x_S, x_T, x_P)||p_{\theta_\alpha}(z_\alpha))$ that models the modality-agnostic information shared across three modalities. While the objective is to learn the shared latent space amongst three modalities, simple approaches like mean-squared error can only be computed between a pair of features. Hence, we first use MAB to combine $\{z_\alpha^S, z_\alpha^T\}$ into z^η and then align it with z_α^P. Additionally, mean-squared error has been shown to be sub-optimal for high-dimensional complex distribution [56] like scene information. Hence, we align the modality-agnostic component from three modalities using InfoNCE [56] to maximising their mutual information between the combined sketch and text z^η and photo z^P (ignoring the subscript α) as,

$$\text{L}_{NCE} = -\mathbb{E}_i\Big[\log \frac{f_{align}(z_i^P, z_i^\eta)}{\sum_j f_{align}(z_j^P, z_i^\eta)}\Big] \tag{11}$$

where, $f_{align}(z^P, z^\eta) = exp((z^P)^T W z^\eta)$. Therefore, the total loss L_{tot} of the unified model for SceneTrilogy consists of the following terms,

$$\begin{aligned} \text{L}_{tot} = \text{L}_{rec} &- \log q(\tau^{-1}(z_{tot}^S|z_\alpha)) - \log|\det J_{\tau^{-1}}(z_{tot}^S|z_\alpha) \\ &- \log q(\tau^{-1}(z_{tot}^T|z_\alpha)) - \log|\det J_{\tau^{-1}}(z_{tot}^T|z_\alpha) \\ &- \log q(\tau^{-1}(z_{tot}^P|z_\alpha)) - \log|\det J_{\tau^{-1}}(z_{tot}^P|z_\alpha) + \text{L}_{NCE} \end{aligned} \tag{12}$$

3.3 Inference

SceneTrilogy for Retrieval: In this section, we show how SceneTrilogy can perform joint sketch- and text-based image retrieval. First, we leverage sketch, text, and image encoders to extract their feature representation as,

$$[z_\alpha^S, z_\beta^S] = \mathcal{F}_{\theta_S}(x_S);\ [z_\alpha^T, z_\beta^T] = \mathcal{F}_{\theta_T}(x_T);\ [z_\alpha^P, z_\beta^P] = \mathcal{F}_{\theta_P}(x_P) \tag{13}$$

The encoded feature representation $z_S \in \mathbb{R}^c$ is split into modality-agnostic $z_\alpha^S \in \mathbb{R}^d$ and modality-specific $z_\beta^S \in \mathbb{R}^{c-d}$. The modality-agnostic component of sketch and text is constructively combined using MAB as,

$$z^\eta = \text{MAB}(\mathcal{V}, rFF(\eta));\ \text{where, } \eta = \{z_\alpha^S, z_\alpha^T\}; \eta \in \mathbb{R}^{2\times d} \tag{14}$$

Finally, fine-grained sketch and text based image retrieval (FG-STBIR) is performed by matching the optimal z_α^P from the gallery that maximise the mutual information $f_{align}(z_\alpha^P, z^\eta)$.

SceneTrilogy for Captioning: We show how Scenetrilogy can simultaneously perform both image and sketch captioning. Given input sketch (x_S) or input photo (x_P), we first extract their feature representations as: $[z^S_\alpha, z^S_\beta] = \mathcal{F}_{\theta_S}(x_S)$ or $[z^P_\alpha, z^P_\beta] = \mathcal{F}_{\theta_P}(x_P)$. Given the modality-agnostic component ($z^S_\alpha \simeq z^P_\alpha \simeq z_\alpha$), we sample from the learned prior distribution $z^T_\beta \sim q(z^T_\beta)$ to predict the total text representation z^T_{tot} using the conditional invertible neural network τ as $z^T_{tot} = \tau(z^T_\beta | z_\alpha)$. Finally, we leverage a text decoder p_φ to generate the predicted text as, $p_\varphi(w_t | z^T_{tot}, w_1, \ldots, w_{t-1})$.

4 Experiments

Datasets: We use two scene sketch dataset with fine-grained alignment between sketch, text, and photo: (a) SketchyCOCO [24] comprise of $14,081$ scene sketches-photo pairs. The photos are taken from the larger MSCOCO dataset [48] comprising of $164K$ scene images with paired textual description. However, most sketches in SketchyCOCO [24] contain less than one foreground instance. Hence, following Liu *et al.* [49], we filter SketchyCOCO to 1015/210 train/test scene sketches. The resulting data contains: scene sketches, images, and image captions. (b) Unlike SketchyCOCO [24], where the scene sketches are synthetically generated, FSCOCO [16] includes $10,000$ (7000/3000 train/test split) human drawn scene sketches with paired textual description of sketches (i.e., sketch captions). The paired photos in FSCOCO are taken from MSCOCO [48]. Hence, the resulting dataset contains: human drawn scene sketches, sketch captions, and photos. In addition, we also evaluate how the proposed unified model generalise to object-level sketches in Song *et al.* [75]. Song *et al.* [75] contains 1374 sketch-text-photo triplets with 1112 and 262 for training and testing respectively.

Implementation Details: Our model is implemented in PyTorch using 11GB Nvidia RTX 2080-Super GPU. First, we pre-train the image encoder and text decoder on the task of image captioning using $82,783$ photo-text pairs (excluding the photos in SketchyCOCO and FSCOCO) for 15 epochs. Next, we jointly train our unified model using SketchyCOCO [24], or FSCOCO [16], or Song *et al.* [75] for 200 epochs. We use Adam optimiser with learning rate 0.0001 and batch size 16. ImageNet pretrained VGG-16 [73] is used as the encoder network for images and sketches with shared parameters ($\mathcal{F}_{\theta_S} \cong \mathcal{F}_{\theta_I}$) followed by a fully connected layer to give an output feature representation $\{z^S_{tot}, z^P_{tot}\} \in \mathbb{R}^{1056}$ dimensions. The resulting feature is split into modality-agnostic component $\{z^S_\alpha, z^P_\alpha\} \in \mathbb{R}^{996}$ and modality-specific feature $\{z^S_\beta, z^P_\beta\} \in \mathbb{R}^{64}$. To encode text, we use a bidirectional GRU unit with 1024 hidden units to give an output latent representation $z^T_{tot} \in \mathbb{R}^{1024}$. We split $z^T_{tot} = [z^T_\alpha, z^T_\beta]$ into text specific feature $z^T_\beta \in \mathbb{R}^{32}$ and a modality agnostic $z^T_\alpha \in \mathbb{R}^{992}$. Following [35], our text decoder is a single layer autoregressive LSTM decoder that predicts the probability distribution over a fixed vocabulary set of $10,010$ words at each time step. For image and sketch decoder, we leverage two separate models using similar architecture of Zhang *et al.* [90] to synthesise sketches/images of 64×64 dimensions. Since our goal is not to generate realistic scene sketches or images, we do not include an additional

discriminator [92] to the output of image or sketch decoder to improve generation quality. Generating realistic output from the decoder requires non-trivial modifications due to the added scene complexity [24]. Following [64],the structure of our conditionally invertible neural network comprise of alternating affine coupling [20], activation normalisation [37], and switch layers [20].

Evaluation Metric: In line with FG-SBIR research, we use Acc.@q accuracy, i.e. percentage of sketches having true matched photo in the top-q list. For sketch/image captioning, we measure accuracy using standard metrics (a higher number is better) BELU (B) 1-4 [57], CIDEr (C) [83], ROUGE (R) [45], METEOR (M) [19], and SPICE [2].

Competitors: We compare against (i) existing state-of-the-art methods that align two modalities (SOTA2): **Triplet-SN** [88] is a FG-SBIR method that employs Sketch-A-Net [89] backbone trained using triplet loss. **HOLEF** [76] extends *Triplet-SN* by adding spatial attention along with higher order ranking loss. **SketchyScene*** is similar to [95] that adopts *Triplet-SN* by replacing the base network from Sketch-A-Net to VGG-16 [73] along with an auxiliary cross-entropy that utilise object category information. **SceneSketcher** [49] performs FG-SBIR using Graph Convolutional Network [39] to model scene sketch layout information. CLIP [61] align text and images using Transformer [72] to encode text and Vision Transformer [21] to encode images. The model is trained on 400 million text-image pairs. We use the publicly available ViT-B/32 weights[3] for **CLIP-zero** (without fine-tuning). Since additional fine-tuning made *CLIP* unstable, we only train the layer normalisation [6] modules in **CLIP**. For captioning, **Show-Attend-Tell** [87] uses an ImageNet pretrained VGG-16 to encode an image followed by an soft-attention mechanism along with an LSTM decoder to predict words at each time step. **GMM-CVAE** [85] employs conditional variational autoencoder with a Gaussian mixture model as the prior distribution to generate diverse captions for image that have multiple interpretation. **AG-CVAE** [85] extends *GMM-CVAE* with a linear combination of Gaussian mixture priors to predict diverse and accurate captions. **LNFMM** [53] is a recent model similar to the proposed method that splits information in each modality into a modality-specific and modality-agnostic information followed by reconstructing text-specific information from the shared component to generate captions. (ii) We compare with models that can align 3 modalities (SOTA3): **QuadSk-Txt** [75] encodes object-level sketch and image using VGG-16 [73] backbone with shared parameters and text using bidirectional LSTM. Extending triplet loss to combine sketch and text for image retrieval using quadruplet loss. **SharedCM** is similar to [5] but uses ResNet-18 [27] with weight sharing from layers *Res-Block-4* onwards to encode each modality into a shared latent space. Retrieval is performed using cosine distance. (iii) In addition to using state-of-the-art models, we design a few baselines: Perceiver [33] is a recent model with ability to encode multiple modalities. However, we found the original architecture to be unstable when trained on small datasets such as SketchyCOCO [24] or FSCOCO [16]. Hence, we design **Set-Atten.** that adopts the cross-attention

[3] https://github.com/openai/CLIP

Table 1. Quantitative results of fine-grained sketch-based image retrieval (FG-SBIR) on two scene sketch datasets [24, 16] and one object-sketch dataset [75].

	Method	Song *et al.* [75]		SketchyCOCO [24]		FSCOCO [16]	
		Acc.@1	Acc.@10	Acc.@1	Acc.@10	Acc.@1	Acc.@10
SOTA2	Triplet-SN [88]	49.2	82.1	6.2	32.9	4.7	21.0
	HOFEL [76]	49.2	83.6	6.2	40.7	4.9	21.7
	SketchyScene* [95]	50.1	83.9	36.5	78.6	23.0	52.3
	SceneSketcher [49]	–	–	31.9	86.2	–	–
SOTA3	QuadSkTxt [75]	50.4	84.7	37.4	87.1	23.6	52.9
	SharedCM [5]	47.5	80.3	37.3	86.8	23.4	52.6
Baseline	CLIP3	48.1	80.8	15.3	43.9	5.5	26.5
	Set-Atten.	50.5	84.9	37.9	87.4	23.7	53.5
	Proposed	50.7	85.1	38.2	87.6	24.1	53.9

Fig. 3. Qualitative results of fine-grained sketch-based image retrieval on FSCOCO [16].

and self-attention mechanism in [33] upon features extracted from VGG-16 (for sketches and images) and bidirectional GRU (for text). **CLIP3** adapts the image encoder in CLIP [61] to support sketch and image encoding using shared parameters of the Visual Transformer [21]. For captioning, we extend *QuadSkTxt* by adding an autoregressive single layer LSTM layer for text decoding in our baseline **QuadSkTxt-D**. Similarly, we adapt *SharedCM* using a LSTM based text decoding module in **SharedCM-D**.

4.1 Fine-Grained Scene Sketch-based Image Retrieval

We perform comparative study on scene sketches from SketchyCOCO [24] and FSCOCO [16]. In addition, to demonstrate generalisation we also evaluate on object-sketches from Song *et al.* [75] dataset. From Fig. 3 and Tab. 1 we make the following observations: (i) SOTA3 slightly outperforms SOTA2 methods due to learning more generalisible features when learning across three modalities instead of two in SOTA2. (ii) *QuadSkTxt* outperforms *SharedCM* indicating sharing model parameters between sketch and image encoder leads to less overfitting and better retrieval performance. (iii) *Set-Atten.* outperforms *CLIP3* since CLIP [61] trained on 400 million images fails to adapt to sketches comprising of a few thousand training samples. (iv) The proposed method outperforms existing methods due to information decoupling/splitting in sketch, text, and photo that filters noisy modality-specific information when mapping sketches with images.

4.2 Fine-Grained Text-based Image Retrieval

While some information are best expressed by drawing, others like colour information are best described via text. In Tab. 2 we compare fine-grained text-based

Table 2. Quantitative results of fine-grained text-based image retrieval (FG-TBIR) on two scene sketch datasets [24, 16] and one object-sketch dataset [75].

Method		Song *et al.* [75]		SketchyCOCO [24]		FSCOCO [16]	
		Acc.@1	Acc.@10	Acc.@1	Acc.@10	Acc.@1	Acc.@10
SOTA2	CLIP-zero [61]	12.8	37.7	21.0	50.9	11.5	35.3
	CLIP [61]	13.1	37.9	22.1	52.3	14.8	36.6
SOTA3	QuadSkTxt [75]	12.6	37.4	11.1	31.1	7.2	23.6
	SharedCM [5]	12.6	37.5	10.7	31.0	6.9	23.1
Baseline	CLIP3	13.1	37.9	22.1	52.3	14.8	36.6
	Set-Atten.	12.8	37.8	20.1	51.0	12.5	35.8
Proposed		13.2	37.9	21.5	51.6	13.7	36.3

Table 3. Quantitative results of fine-grained sketch and text based image retrieval (FG-STBIR) on two scene sketch datasets [24, 16] and one object-sketch dataset [75].

Method		Song *et al.* [75]		SketchyCOCO [24]		FSCOCO [16]	
		Acc.@1	Acc.@10	Acc.@1	Acc.@10	Acc.@1	Acc.@10
SOTA3	QuadSkTxt [75]	52.7	87.0	38.9	87.9	25.1	54.5
	SharedCM [5]	52.1	86.6	38.5	87.3	24.3	54.1
Baseline	CLIP3	49.1	82.0	24.0	53.7	15.9	38.5
	Set-Atten.	52.4	86.8	39.1	88.2	25.3	54.8
Proposed		53.1	87.9	39.5	88.7	25.7	55.2

image retrieval (FG-TBIR) and make the following observations: (i) Given the same amount of train/test split, sketches outperform text as a query modality for fine-grained image retrieval. (ii) *CLIP* outperforms SOTA3 methods due to superior pre-trained weights using 400 million text-image pairs. (iii) The proposed method outperforms most existing methods due to information splitting via conditional invertible neural network. In particular, the proposed method is competitive with *CLIP* in spite of being trained on a few thousand samples.

4.3 Fine-grained Sketch and Text based Image Retrieval

Unlike existing methods that either requires a user to input a query in sketch [88] or text [61] modalities, SceneTrilogy provides the flexibility to a user to provide input in either sketch, or text, or both. While prior studies [75] have shown object-sketch and text have a complimentary nature, we investigate this complimentary nature for scene-level sketches and text. From Tab. 3 we draw the following observations: (i) Combining sketch and text leads to superior fine-grained sketch and text based image retrieval (FG-STBIR) than either of FG-SBIR or FG-TBIR. This shows that scene-level sketches and text are complimentary in nature as combining both these query modalities improves retrieval performance. (ii) *CLIP3* outperforms SOTA3 methods due to superior pre-trained weights. (iii) *Set-Atten.* outperforms *CLIP3* since the latter was pre-trained on large-scale text-image pairs fails to adapt to scene-sketches using just a few thousand training samples. (iv) The proposed method outperforms existing approaches due to information splitting to remove noisy modality-specific information, better modeling of information across sketch and text using multihead attention, and use of InfoNCE [56] instead of naive mean-squared distance.

Table 4. Quantitative results of image captioning on MSCOCO [48] dataset. BL represents Baseline methods. We present the evaluation metrics in the oracle setting i.e., taking the maximum score for each accuracy metric over 100 candidate captions, consistent with previous [53, 85] methods.

	Method	B-1	B-2	B-3	B-4	M	R	C	S
SOTA2	Show-Attend-Tell [87]	71.8	50.4	35.7	25.0	23.0	–	–	–
	GMM-CVAE [85]	86.5	74.0	62.5	52.7	32.9	67.0	143.0	26.3
	AG-CVAE [85]	88.3	76.7	65.4	55.7	34.5	69.0	151.7	27.7
	LNFMM [53]	92.0	80.2	69.5	59.7	40.2	72.9	170.5	31.6
BL	QuadSkTxt-D	71.8	50.7	35.7	25.1	23.2	54.7	95.3	15.7
	SharedCM-D	72.3	51.0	35.7	25.5	23.2	54.9	95.1	15.9
	Proposed	93.1	80.1	70.9	60.3	40.5	72.3	190.2	31.9

4.4 Image Captioning

Following existing literature [85, 53], we present the upper-bound Oracle performance by taking the maximum score for each metric (BELU, CIDEr, ROUGE, METEOR, SPICE) over all candidate captions generated by 100 samples of text-specific information z_β^T from a given modality agnostic information z_α. For *Show-Attend-Tell*, we generate candidate captions by performing a beam search with width 100. For *GMM-CVAE*, *AG-CVAE*, and *LNFMM* we sample the latent space z 100 times to generate candidate captions. From Tab. 4 we observe that (i) naive baselines such as *QuadSkTxt-D*, *SharedCM-D* perform significantly lower than VAE-based existing SOTA2 methods. (ii) Using cINN to model prior distribution in *LNFMM* leads to superior performance than using simple Gaussian priors in *GMM-CVAE* and *AG-CVAE*. (ii) The *Proposed* method is competitive with SOTA2 due to it ability to model complex priors and splitting information during transfer and ability to regenerate the missing text-specific information given the shared modality agnostic component.

4.5 Sketch Captioning

Sketches can better capture saliency information as compared to images. Hence, a sketch captioning module essentially describes the salient/relevant regions of a scene. In addition, unlike text, sketch is a universal language for communication [65]. This enables potential applications of sketch captioning modules in overcoming language barrier across generations or geographical regions. We evaluate sketch captioning using the same metrics popularly used in image captioning literature. From Fig. 4 and Tab. 5 we conclude: (i) LNFMM [53] performs best among SOTA2 methods that signify the importance of information splitting into modality-agnostic and modality-specific components followed by generating text-specific information from the modality-agnostic component from sketch. (ii) The *Proposed* method outperforms both SOTA2 and Baselines due to its ability to handle drastic gap in sketch and image modalities using information split to extract the transferable shared information and reconstruction of modality-specific information from the shared component.

Fig. 4. Qualitative results of candidate caption generated for sketch captioning on [16].

Table 5. Quantitative results of sketch captioning on FSCOCO [16] dataset. BL represents Baseline methods. We present the evaluation metrics in the oracle setting i.e., taking the maximum score for each accuracy metric over 100 candidate captions, consistent with previous [53, 85] methods.

	Method	B-1	B-2	B-3	B-4	M	R	C	S
SOTA2	Show-Attend-Tell [87]	46.2	29.1	17.8	13.7	17.1	44.9	69.4	14.5
	GMM-CVAE [85]	49.6	33.9	18.2	15.5	18.3	48.7	77.6	15.5
	AG-CVAE [85]	50.9	34.1	19.2	16.0	18.9	49.1	80.5	15.8
	LNFMM [53]	52.2	35.7	20.0	16.7	21.0	52.9	90.1	16.0
BL	QuadSkTxt-D	53.7	34.7	21.9	17.9	20.6	52.9	89.3	16.2
	SharedCM-D	53.8	34.7	22.0	18.1	20.6	53.5	89.7	16.2
	Proposed	56.9	39.4	24.8	19.3	21.6	56.6	106.5	18.9

4.6 Ablation

We evaluate the contribution of the key design choices in our proposed method in Tab. 6 as: (i) Replacing InfoNCE [56] with mean-square error drops performance by 0.2/0.3/0.6/0.3/0.6 for Acc.@1/Acc.@10/B-1/B-4/C in FSCOCO [16]. (ii) Additionally, replacing Multihead Attention (MAB) with quadruplet loss [75] leads to a performance drop by 0.6/0.6/0.6/0.3/0.6. (iii) Finally, replacing conditional invertible neural network (cINN) with a simple VAE-like Gaussian prior [68] results in a significant performance drop of 0.6/0.6/3.2/1.4/17.2.

Table 6. Ablation study on FG-STBIR and Sketch Captioning using FSCOCO [16].

cINN	MAB	InfoNCE	Acc.@1	Acc.@10	B-1	B-4	C
✗	✗	✗	25.1	54.6	53.7	17.9	89.3
✓	✗	✗	25.1	54.6	56.3	19.0	105.9
✓	✓	✗	25.5	54.9	56.3	19.0	105.9
✓	✓	✓	25.7	55.2	56.9	19.3	106.5

5 Conclusion

We have studied for the first time the trilogy of relationship among scene-level sketch, text, and photo by introducing scene-sketch in the context of scene understanding. We proposed a unified framework to jointly model sketch, text, and photo that seamlessly supports several downstream task such as fine-grained sketch-based image retrieval, fine-grained text-based image retrieval, fine-grained sketch and text based image retrieval, sketch captioning. We hope future research will further explore challenging non-trivial downstream tasks such as scene-level sketch-based image generation, sketch and text based image generation, and text-based sketch generation tasks.

References

1. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. In: ICLR (2017)
2. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: semantic propositional image caption evaluation. In: ECCV (2016)
3. Aneja, J., Agrawal, H., Batra, D., Schwing, A.G.: Sequential latent space for modeling the intention during image captioning. In: ICCV (2019)
4. Ardizzone, L., Lüth, C., Kruse, J., Rother, C., Köthe, U.: Guided image generation with conditional invertible neural networks. arXiv preprint arXiv:1907.02392 (2019)
5. Aytar, Y., Castrejon, L., Vondrick, C., Pirsiavash, H., Torralba, A.: Cross-modal scene networks. TPAMI (2018)
6. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
7. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: data2vec: A general framework for self-supervised learning in speech, vision and language. arXiv preprint arXiv:2202.03555 (2022)
8. Bhunia, A.K., Chowdhury, P.N., Yang, Y., Hospedales, T.M., Xiang, T., Song, Y.Z.: Vectorization and rasterization: Self-supervised learning for sketch and handwriting. In: CVPR (2021)
9. Bhunia, A.K., Das, A., Riaz Muhammad, U., Yang, Y., Hospedales, T.M., Xiang, T., Gryaditskaya, Y., Song, Y.Z.: Pixelor: A competitive sketching ai agent. so you think you can beat me? In: SIGGRAPH Asia (2020)
10. Bhunia, A.K., Yang, Y., Hospedales, T.M., Xiang, T., Song, Y.Z.: Sketch less for more: On-the-fly fine-grained sketch based image retrieval. In: CVPR (2020)
11. Castrejón, L., Aytar, Y., Vondrick, C., Pirsiavash, H., Torralba, A.: Learning aligned cross-modal representations from weakly aligned data. In: CVPR (2016)
12. Chen, S.Y., Su, W., Gao, L., Xia, S., Fu, H.: Deepfacedrawing: Deep generation of face images from sketches. ACM TOG (2020)
13. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: interpretable representation learning by information maximizing generative adversarial nets. In: NeurIPS (2016)
14. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: ECCV (2020)
15. Cheung, B., Livezey, J.A., Bansal, A.K., Olshaussen, B.A.: Discovering hidden factors of variation in deep networks. arXiv preprint arXiv:2203.02013 (2014)
16. Chowdhury, P.N., Sain, A., Gryaditskaya, Y., Bhunia, A.K., Xiang, T., Song, Y.Z.: Fs-coco: Towards understanding of freehand sketches of common objects in context. arXiv preprint arXiv:2203.02113 (2022)
17. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: CVPR (2020)
18. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
19. Denkowski, M.J., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: WMT@ACL (2014)
20. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estiamtion using real nvp. In: ICLR (2017)
21. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.:

An image is wort 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
22. Du, Y., Li, S., Sharma, Y., Tenenbaum, J.B., Mordatch, I.: Unsupervised learning of compositional energy concepts. In: NeurIPS (2021)
23. Eitz, M., Hays, J., Alexa, M.: How do humans sketch objects? TOG (2012)
24. Gao, C., Liu, Q., Wang, L., Liu, J., Zou, C.: Sketchycoco: Image generation from freehand scene sketches. In: CVPR (2020)
25. Ge, S., Goswami, V., Zitnick, C.L., Parikh, D.: Creative sketch generation. In: ICLR (2021)
26. Gómez, L., Mafla, A., Rusiñol, M., Karatzas, D.: Single shot scene text retrieval. In: ECCV (2018)
27. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
28. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinivk, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: ICLR (2017)
29. Hoffman, M.D., Johnson, M.J.: Elbo surgery: yet another way to carve up the variational evidence lower bound. In: Symposium on Advances in Approximate Baesian Inference, NeurIPS (2016)
30. Hu, C., Li, D., Song, Y.Z., Xiang, T., Hospedales, T.M.: Sketch-a-classifier: Sketch-based photo classifier generation. In: CVPR (2018)
31. Huang, H., Jai, V., Mehta, H., Ku, A., Magalhaes, G., Baldridge, J., Ie, E.: Transferable representation learning in vision-and-language navigation. In: ICCV (2019)
32. Huang, Z., Fu, H., Lau, R.W.H.: Data-driven segmentation and labeling of freehand sketches. ACM TOG (2014)
33. Jargle, A., Gimeno, F., Brock, A., Zisserman, A., Vinyals, O., Carreira, J.: Perceiver: General perception with iterative attention. In: ICML (2021)
34. Karaletsos, T., Belongie, S., Rätsch, G.: Bayesian representation learning with oracle constraints. In: ICLR (2016)
35. Karpathy, A., Li, F.F.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
36. Kim, H., Mnih, A.: Disentangling by factorising. In: ICML (2018)
37. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. In: NeurIPS (2018)
38. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: NeurIPS (2014)
39. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017)
40. Lee, J., Lee, Y., Kim, J., Kosiorek, A.R., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. In: ICML (2019)
41. Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D., Zhou, M.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: AAAI (2020)
42. Li, L., Zou, C., Zheng, Y., Su, Q., Fu, H., Tai, C.L.: Sketch-r2cnn: An attentive network for vector sketch recognition. arXiv preprint arXiv:1811.08170 (2018)
43. Li, T.M., Lukáč, M., Michaël, G., Ragan-Kelley, J.: Differentiable vector graphics rasterization for editing and learning. TOG (Proc. SIGGRAPH Asia) (2020)
44. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., Gao, J.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: ECCV (2020)
45. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out (2004)

46. Lin, H., Fu, Y., Jiang, Y.G., Xue, X.: Sketch-bert: Learning sketch bidirectional encoder representation from transformers by self-supervised learning of sketch gestalt. In: CVPR (2020)
47. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
48. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: common objects in context. In: ECCV (2014)
49. Liu, F., Zhou, C., Deng, X., Zuo, R., Lai, Y.K., Ma, C., Liu, Y.J., Wang, H.: Scenesketcher: Fine-grained image retrieval with scene sketches. In: ECCV (2020)
50. Liu, K., Li, Y., Xu, N., Natarajan, P.: Learn to combine modalities in multimodal deep learning. arXiv preprint arXiv:1805.11730 (2018)
51. Liu, X., Li, H., Shao, J., Chen, D., Wang, X.: Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In: ECCV (2018)
52. Lyu, Y., Liang, P.P., Deng, Z., Salakhutdinov, R., Morency, L.P.: Dime: Fine-grained interpretations of multimodal models via disentangled local explanations. arXiv preprint arXiv:2203.02013 (2022)
53. Mahajan, S., Gurevych, I., Roth, S.: Latent normalizing flows for many-to-many cross-domain mappings. In: ICLR (2020)
54. Mahajan, S., Roth, S.: Diverse image captioning with context-object split latent spaces. In: NeurIPS (2020)
55. Mo, H., Edgar, S.S., Gao, C., Zou, C., Wang, R.: General virtual sketching framework for vector line art. TOG (2021)
56. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
57. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002)
58. Portenier, T., Hu, Q., Szabó, A., Bigdeli, S.A., Favaro, P., Zwicker, M.: Faceshop: Deep sketch-based face image editing. ACM TOG (2018)
59. Qi, Y., Su, G., Chowdhury, P.N., Li, M., Song, Y.Z.: Sketchlattice: Latticed representation for sketch manipulation. In: ICCV (2021)
60. Qian, Y., Yang, Y., Song, Y.Z., Xiang, T., Hospedales, T.: Sketch-a-net that beats humans. In: BMVC (2015)
61. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
62. Reed, S., Sohn, K., Zhang, Y., Lee, H.: Learning to disentangle factors of variation with manifold interaction. In: ICML (2014)
63. Riberio, L.S.F., Bui, T., Collomosse, J., Ponti, M.: Scene designer: a unified model for scene search and synthesis from sketch. In: Workshop on Sketching for Human Expressivity, ICCV (2021)
64. Rombach, R., Esser, P., Ommer, B.: Network-to-network translation with conditional invertible neural networks. In: NeurIPS (2020)
65. Rossi, P.: Logic and the Art of Memory: The Quest for a Universal Language. University of Chicago Press (2001)
66. Rubenstein, P.K., Schoelkopf, B., Tolstikhin, I.: Wasserstein auto-encoders: Latent dimensionality and random encoders. In: ICLR Workshop (2018)
67. Sain, A., Bhunia, A.K., Yang, Y., Xiang, T., Song, Y.Z.: Cross-modal hierarchical modelling for fine-grained sketch based image retrieval. In: BMVC (2020)
68. Sain, A., Bhunia, A.K., Yang, Y., Xiang, T., Song, Y.Z.: Stylemeup: Towards style-agnostic sketch-based image retrieval. In: CVPR (2021)

69. Sarvadevabhatla, R.K., Dwivedi, I., Biswas, A., Manocha, S., R., V.B.: Sketchparse: Towards rich descriptions for poorly drawn sketches using multi-task hierarchical deep networks. In: ACM MM (2017)
70. Schneider, R.G., Tuytelaars, T.: Sketch classification and classification-driven analysis using fisher vectors. ACM TOG (2014)
71. Schneider, R.G., Tuytelaars, T.: Example-based sketch segmentation and labeling using crfs. ACM TOG (2016)
72. Shazeer, V.A., Parmar, N., Uszkoreit, N., Jones, J., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
73. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
74. Song, J., Ermon, S.: Multi-label contrastive predictive coding. In: NeurIPS (2020)
75. Song, J., Song, Y.Z., Xiang, T., Hospedales, T.: Fine-grained image retrieval: the text/sketch input dilemma. In: BMVC (2017)
76. Song, J., Yu, Q., Song, Y.Z., Xiang, T., Hospedales, T.M.: Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In: ICCV (2017)
77. Su, G., Qi, Y., Pang, K., Yang, J., Song, Y.Z.: Sketchhealer: A graph-to-sequence network for recreating partial human sketches. In: BMVC (2020)
78. Sun, Z., Wang, C., Zhang, L., Zhang, L.: Free hand-drawn sketch segmentation. In: ECCV (2012)
79. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: ICLR (2020)
80. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. In: 37th annual Allerton Conf. on Communication, Control, and Computing (1999)
81. Tishby, N., Zaslavsky, N.: Deep learning and the information bottleneck principle. In: Information Theory Workshop (ITW) (2015)
82. Tsai, Y.H.H., Liang, P.P., Zadeh, A., Morency, L.P., Salakhutdinov, R.: Learning factorized multimodal representations. In: ICLR (2019)
83. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR (2015)
84. Vinker, Y., Pajouheshgar, E., Bo, J.Y., Bachmann, R.C., Bermano, A.H., Cohen-Or, D., Zamir, A., Shamir, A.: Clipasso: Semantically-aware object sketching. arXiv preprint arXiv:2202.05822 (2022)
85. Wang, L., Schwing, A.G., Lazebnik, S.: Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In: NeurIPS (2017)
86. Wang, Z., Liu, X., Li, H., Sheng, L., Yan, J., Wang, X., Shao, J.: Camp: Cross-modal adaptive message passing for text-image retrieval. In: ICCV (2019)
87. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML (2015)
88. Yu, Q., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.M., Loy, C.C.: Sketch me that shoe. In: CVPR (2016)
89. Yu, Q., Yang, Y., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.M.: Sketch-a-net: A deep neural network that beats humans. IJCV (2017)
90. Zhang, H., Goodfellow, I.J., Metaxas, D.N., Odena, A.: Self-attention generative adversarial networks. In: ICML (2019)
91. Zhang, H., Koh, J.Y., Baldrige, J., Lee, H., Yang, Y.: Cross-modal contrastive learning for text-to-image generation. arXiv preprint arXiv:2101.04702 (2021)

92. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan++: Realistic image synthesis with stacked generative adversarial networks. In: TPAMI (2019)
93. Zhou, B., Lapedriza, A., Khosla, A., Olivia, A., Torralba, A.: Places: A 10 million image database for scene recognition. TPAMI (2017)
94. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. IJCV (2019)
95. Zou, C., Yu, Q., Du, R., Mo, H., Song, Y.Z., Xiang, T., Gao, C., Chen, B., Zhang, H.: Sketchyscene: Richly-annotated scene sketches. In: ECCV (2018)

Supplemental for SceneTrilogy: On Scene Sketches and its Relationship with Text and Photo

Pinaki Nath Chowdhury[1,2] Ayan Kumar Bhunia[1]
Tao Xiang[1,2] Yi-Zhe Song[1,2]
[1]SketchX, CVSSP, University of Surrey, United Kingdom.
[2]iFlyTek-Surrey Joint Research Centre on Artificial Intelligence.
{p.chowdhury, a.bhunia, t.xiang, y.song}@surrey.ac.uk

A Additional Related Work

Disentangled Representation: Learning disentangled representation means decomposing a feature representation into distinct informative factors such that each factor can explain a particular variation of the data. One of the key benefits of disentangled representation is its ability to improve both generative and discriminative performance in multimodal tasks [82, 52]. Current literature in disentangled representation can be broadly divided into two categories: (a) the factors of variations are known where each variation is individually controlled using supervised training [15, 34, 62], and (b) the factors of variations are partially known or unknown [68, 28, 38, 66] where learning occurs in an unsupervised setting. Our proposed framework aligns more closely to the partially known or unknown setup. In particular, a similar method to ours is InfoGAN [13] that learns disentangled representations by proposing an information-theoric regularization to maximises the mutual information between the latent variable and the generated distribution. However, due to the instability of GANs, later works primarily focused on enforcing a factorial distribution [36] of representation using modification on VAE [38, 28]. Inspired by the benefits of disentangled representation on generative [13, 28] and retrieval tasks [68], especially for learning multi-modal representations [82], recent works achieve this disentanglement using flow-based conditional invertible neural networks [53, 64]. However, such works have typically been restricted to the simpler setting of two modalities (text and photo). We extend this to a complex setup by aligning three heterogeneous modalities (sketch, text, and photo) together under the umbrella of SceneTrilogy.

B Limitations and Future Work

The central objective of SceneTrilogy is to achieve *many-to-many* mapping across scene-level sketch, text, and photo. This includes additional tasks that our unified framework currently fails to address like (i) sketch to photo generation [24, 63] and (ii) photo to sketch generation [55, 84]. Existing models achieve sketch to photo generation by leveraging a two-step process where the first stage involves modeling object-level information followed by a second stage to model scene-level information. Our *single-step* unified model fails to implicitly handle object-

and scene-level hierarchical information. A promising direction for future work could be to implicitly learn hierarchical information by utilising compositional energy concepts [22]. Extending SceneTrilogy for image to sketch generation is relatively simpler with the use of differentiable rasterisation [43]. However, integrating sketch to photo or, photo to sketch in SceneTrilogy is non-trivial. Hence we leave these tasks as future works.

C Derivations

In this section, we derive the joint distribution of SceneTrilogy represented as:

$$\begin{aligned}\log p_\theta(x_S, x_T, x_P) \geq & \mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log p_\theta(x_S, x_T, x_P|z)] \\ & \mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log p_\theta(z) - \log q_\phi(z|x_S, x_T, x_P)]\end{aligned} \tag{15}$$

We begin by understanding the variational lower bound on marginal likelihood defined in [38, 29] as:

$$\begin{aligned}&\log p_\theta(x) = \int p_\theta(x, z)dz = \int q_\phi(z|x)\frac{p_\theta(x, z)}{q_\phi(z|x)}dz \geq \mathbb{E}_{q_\phi(z|x)} \log \frac{p_\theta(x, z)}{q_\phi(z|x)} \\ &\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z) + \log p_\theta(z)] - \mathbb{E}_{q_\phi(z|x)} \log q_\phi(z|x)\end{aligned} \tag{16}$$

Extending to three variables i.e., sketch (x_S), text (x_T), and photo (x_P) we get,

$$\begin{aligned}\log p_\theta(x_S, x_T, x_P) \geq & \mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log p_\theta(x_S, x_T, x_P|z)] \\ & + \mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log p_\theta(z) - \log q_\phi(z|x_S, x_T, x_P)]\end{aligned} \tag{17}$$

The latent variable z comprises of three components $z = \{z^S_{tot}, z^T_{tot}, z^P_{tot}\}$ each representing the latent variable for sketch, text, and photo. As shown in Figure 2, the latent variable for each modality representing the intra-modality information can be decoupled into a modality-agnostic and modality-specific components as: $z^S_{tot} = [z^S_\alpha, z^S_\beta]$; $z^T_{tot} = [z^T_\alpha, z^T_\beta]$; and $z^P_{tot} = [z^P_\alpha, z^P_\beta]$. Considering a high mutual information among the modality-agnostic components such that $z^S_\alpha \simeq z^T_\alpha \simeq z^P_\alpha \simeq z_\alpha$, we factorise each term in Equation 17 as,

$$\begin{aligned}&\mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log p_\theta(x_S, x_T, x_P|z)] = \\ &\qquad \mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log p_\theta(x_S, x_T, x_P|z_\alpha, z^S_\beta, z^T_\beta, z^P_\beta)]\end{aligned} \tag{18}$$

Given the modality-specific $\{z^S_\beta, z^T_\beta, z^P_\beta\}$ and the modality-agnostic z_α, we can reconstruct their respective data distribution $p(x_S)$, $p(x_T)$, $p(x_P)$. Hence we can re-write Equation 18 as,

$$\begin{aligned}&\mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log p_\theta(x_S, x_T, x_P|z_\alpha, z^S_\beta, z^T_\beta, z^P_\beta) \simeq \\ &\qquad \mathbb{E}_{q_{\phi_1}(z_\alpha|x_S,x_T,x_P)q_{\phi_2}(z^S_\beta|x_S,z_\alpha)}[\log p_\theta(x_S|z_\alpha, z^S_\beta)] \\ &\qquad + \mathbb{E}_{q_{\phi_1}(z_\alpha|x_S,x_T,x_P)q_{\phi_3}(z^T_\beta|x_T,z_\alpha)}[\log p_\theta(x_T|z_\alpha, z^T_\beta)] \\ &\qquad + \mathbb{E}_{q_{\phi_1}(z_\alpha|x_S,x_T,x_P)q_{\phi_4}(z^P_\beta|x_P,z_\alpha)}[\log p_\theta(x_P|z_\alpha, z^P_\beta)]\end{aligned} \tag{19}$$

Each term analogous to $p_\theta(x_S|z_\alpha, z_\beta^S)$ is known as the reconstruction terms and is modeled using a modality-specific decoder network as described in Section 4. Next, we factorise the prior $\log p_\theta(z)$ in Equation 17 as,

$$\begin{aligned}\mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log p_\theta(z)] &= \mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log p_\theta(z_\alpha, z_\beta^S, z_\beta^T, z_\beta^P)] \\ &= \mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log p_\theta(z_\beta^S, z_\beta^T, z_\beta^P|z_\alpha) + \log p_\theta(z_\alpha)]\end{aligned} \tag{20}$$

Considering a low mutual information between the modality-specific components such that $p(z_\beta^S, z_\beta^T, z_\beta^P) \simeq p(z_\beta^S) \cdot p(z_\beta^T) \cdot p(z_\beta^P)$ from Equation 20 we get,

$$\begin{aligned}&\mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log p_\theta(z_\beta^S, z_\beta^T, z_\beta^P|z_\alpha) + \log p_\theta(z_\alpha)] \simeq \\ &\mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log p_{\theta_S}(z_\beta^S|z_\alpha) + \log p_{\theta_T}(z_\beta^T|z_\alpha) + \log p_{\theta_P}(z_\beta^P|z_\alpha) + \log p_{\theta_\alpha}(z_\alpha)]\end{aligned} \tag{21}$$

Finally, we factorise the variational posterior $q_\phi(z|x_S, x_T, x_P)$ in Equation 17 as,

$$\begin{aligned}&\mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log q_\phi(z_\alpha, z_\beta^S, z_\beta^T, z_\beta^P|x_S, x_T, x_P)] \simeq \\ &\quad\mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log q_\phi(z_\beta^S|x_S, x_T, x_P, z_\alpha) + \log q_\phi(z_\beta^T|x_S, x_T, x_P, z_\alpha) \\ &\qquad + \log q_\phi(z_\beta^P|x_S, x_T, x_P, z_\alpha) + \log q_\phi(z_\alpha|x_S, x_T, x_P)]\end{aligned} \tag{22}$$

Since modality-specific information (say z_β^S) is independent of other modalities (text x_T and photo x_P), we can simplify Equation 22 as,

$$\begin{aligned}&\mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log q_\phi(z_\alpha, z_\beta^S, z_\beta^T, z_\beta^P|x_S, x_T, x_P)] \simeq \\ &\quad\mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log q_{\phi_\alpha}(z_\alpha|x_S, x_T, x_P) \\ &\qquad + \log q_{\phi_S}(z_\beta^S|x_S, z_\alpha) + \log q_{\phi_T}(z_\beta^T|x_T, z_\alpha) + \log q_{\phi_P}(z_\beta^P|x_P, z_\alpha)]\end{aligned} \tag{23}$$

Finally, combining Equation 17, 21, and 23 we get our desired KL-divergence terms in Equation 4 as,

$$\begin{aligned}&\mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log q_\phi(z|x_S, x_T, x_P) - \log p_\theta(z)] = \\ &\qquad\mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log q_{\phi_\alpha}(z_\alpha|x_S, x_T, x_P) - \log p_{\theta_\alpha}(z_\alpha)] \\ &\qquad + \mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log q_{\phi_S}(z_\beta^S|x_S, z_\alpha) - \log p_{\theta_S}(z_\beta^S|z_\alpha)] \\ &\qquad + \mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log q_{\phi_T}(z_\beta^T|x_T, z_\alpha) - \log p_{\theta_T}(z_\beta^T|z_\alpha)] \\ &\qquad + \mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log q_{\phi_P}(z_\beta^P|x_P, z_\alpha) - \log p_{\theta_P}(z_\beta^P|z_\alpha)]\end{aligned} \tag{24}$$

Using the definition of KL-divergence as $D_{KL}(P||Q) = \mathbb{E}_{P(x)} \log \frac{P(x)}{Q(x)}$ we get,

$$\begin{aligned}&\mathbb{E}_{q_\phi(z|x_S,x_T,x_P)}[\log q_\phi(z|x_S, x_T, x_P) - \log p_\theta(z)] = \\ &D_{KL}(q_{\phi_\alpha}(z_\alpha|x_S, x_T, x_P)||p_{\theta_\alpha}(z_\alpha)) + D_{KL}(q_{\phi_S}(z_\beta^S|x_S, z_\alpha)||p_{\theta_S}(z_\beta^S|z_\alpha) \\ &+ D_{KL}(q_{\phi_T}(z_\beta^T|x_T, z_\alpha)||p_{\theta_T}(z_\beta^T|z_\alpha) + D_{KL}(q_{\phi_P}(z_\beta^P|x_P, z_\alpha)||p_{\theta_P}(z_\beta^P|z_\alpha)\end{aligned} \tag{25}$$

D Additional Discussions

What is represented in the modality-specific component?
We learn to decouple modality-specific and modality-agnostic in an unsupervised setup, similar to [68,28,13]. A limitation of this learning setup is the lack of interpretability [68] where the relevance of each learned factor is uncontrollable. While an alternative is to learn disentangled representations in a supervised setting, there are several limitations that make it an undesired learning paradigm [36] such as, (i) humans can omit factors or have inconsistency in labels assigned for variations that are difficult to identify (ii) obtaining labels is costly that requires a human in the loop (iii) humans are capable of learning factors of variations in an unsupervised setting. Although learning *exactly* what is learned in the modality-specific components is admittedly difficult, Section 4 reveals some intuitions. From Table 6, we observe that replacing conditional invertible neural network with a sub-optimal Gaussian prior leads to a greater drop in image captioning metrics (3.2/1.4/17.2 in B-1/B-4/C) than retrieval metrics (0.6/0.6 in Acc.@1/Acc.@10). This indicates that learning a *good* text-specific information is more essential for the task of text generation (i.e., constructing sentences using grammatical knowledge given the semantic information). Additionally, existing literature [68] has shown that sketch-specific information primarily captures drawing styles. We hope future research would further make the *black-box* nature of unsupervised learning of disentangled features interpretable.

Why not use multi-modal pretraining using Data2Vec?
Data2Vec [7] is a generalised approach for self-supervised learning across different modalities. While a similar technique could be incorporated in SceneTrilogy as a pre-training step, such improvements is orthogonal to our current objective of establishing SceneTrilogy with our proposed unified framework being the first stab in this direction.

Why not use CMPlaces dataset?
While Castrejón *et al.* [11] studied a cross-modal problem of 5 modalities that includes sketch, text, and photo, it crucially lacks fine-grained pairings togethers i.e., each modality is sourced independently without referencing to another. Hence, it is incompatible with our setting where instance-level pairings are necessitated to study differences in scene representations amongst various modalities.

Why not use SOTA (especially for TBIR) methods like CAMP, Unicoder-VL, Uniter, Oscar?
SceneTrilogy aims to learn the relation among scene-level sketch, text, and photo. While methods like CAMP [86], Unicoder-VL [41], Uniter [14], and Oscar [44] requires millions of paired text and photo, existing sketch datasets are limited to a few thousands. More importantly, two of the dataset (SketchyCOCO [24] and FSCOCO [16]) used to study SceneTrilogy, are derived from MSCOCO [48] (i.e., the training and testing set of [24] and [16] are from MSCOCO). However, SOTA methods like [86,41,14,44] used the entire MS-COCO for training. Hence, comparing [86,41,14,44] on datasets that are subsets of MS-COCO [48] is unfair.