SceneTrilogy: On Human Scene-Sketch and its Complementarity with Photo and Text

Pinaki Nath Chowdhury[1,2] Ayan Kumar Bhunia[1] Aneeshan Sain[1,2] Subhadeep Koley[1,2]
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, a.sain, s.koley, t.xiang, y.song}@surrey.ac.uk

Abstract

In this paper, we extend scene understanding to include that of human sketch. The result is a complete trilogy of scene representation from three diverse and complementary modalities – sketch, photo, and text. Instead of learning a rigid three-way embedding and be done with it, we focus on learning a flexible joint embedding that fully supports the "optionality" that this complementarity brings. Our embedding supports optionality on two axes: (i) optionality across modalities – use any combination of modalities as query for downstream tasks like retrieval, (ii) optionality across tasks – simultaneously utilising the embedding for either discriminative (e.g., retrieval) or generative tasks (e.g., captioning). This provides flexibility to end-users by exploiting the best of each modality, therefore serving the very purpose behind our proposal of a trilogy in the first place. First, a combination of information-bottleneck and conditional invertible neural networks disentangle the modality-specific component from modality-agnostic in sketch, photo, and text. Second, the modality-agnostic instances from sketch, photo, and text are synergised using a modified cross-attention. Once learned, we show our embedding can accommodate a multi-facet of scene-related tasks, including those enabled for the first time by the inclusion of sketch, all without any task-specific modifications. Project Page: http://www.pinakinathc.me/scenetrilogy

1. Introduction

Scene understanding sits at the very core of computer vision. As object-level research matures [25, 33], an encouraging shift can be observed in recent years on scene-level tasks, e.g., scene recognition [118], scene captioning [56], scene synthesis [35], and scene retrieval [14, 58].

Scene research has generally progressed from that of single modality [118, 119] to the very recent focus on multi-modality [3, 14, 20]. The latter setting not only triggered a series of practical applications [35, 58, 106, 120] 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 [60, 62, 63], via applications such as text-based scene retrieval (TBIR) [36], and scene captioning [24, 62, 63].

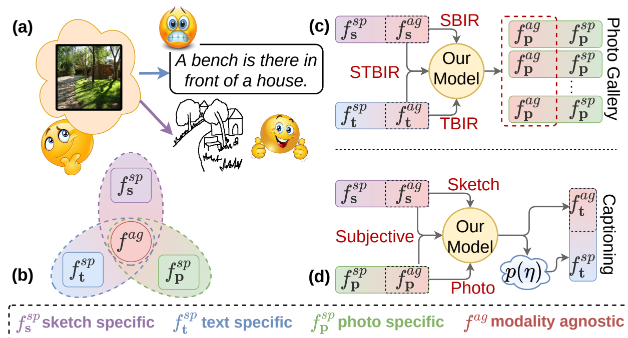


Figure 1. Some scenes are easy to describe via sketch; for others, text is better. We provide the option to sketch, write, or both (sketch+text). For "optionality" across tasks, we disentangle sketch, text, and photo into a discriminative (e.g., retrieval) part f^{ag} shared across modalities, and a generative (e.g., captioning) part specific to one modality (f_s^{sp}, f_t^{sp} f_p^{sp}). This supports a multi-facet of scene-related tasks without task-specific modifications.

This paper follows the said trend of multi-modal scene understanding and extend it to also include human scene-sketch. Sketch is identified because of its unique characteristics of being both expressive and subjective, evident in an abundance of object-level sketch research [12], and very recently on scene-level [20]. To verify there is indeed useful complementarity that sketch can bring to multi-modal scene understanding, we first conducted two pilot studies (i) on expressivity, we compare text and sketch in terms of scene image retrieval, and (ii) on subjectivity, we test a novel task of subjective captioning where sketch or parts-of-speech [27] are used as guidance for image captioning. On (i), results show there is significant disagreement in terms of retrieval accuracy when one is used as query over the other, indicating there is complementary information between the two modalities. On (ii), sketch is shown to offer more subjectivity as a guiding signal than text, when quantified using common metrics such as BELU-4 [69] and CIDEr [99].

To fully explore the complementarity of all three modalities, we desire a flexible joint embedding that best sustains "optionality" across *modalities*, and also across *tasks*. The former enables end-users to use any combination of modalities (e.g., only sketch, only text, or both sketch+text) as a query for downstream tasks; and the latter provides option of utilising the learned embedding for both discriminative (e.g., retrieval) and generative problems (e.g., captioning).

This desired level of "optionality" is however not achievable via naive three-way joint embeddings common in the literature [3, 14, 20]. Instead, we advocate a three-way disentanglement (Fig. 1(b)), where each of the three modalities is disentangled into their modality-specific component ($f_{\mathbf{s}}^{sp}$, $f_{\mathbf{p}}^{sp}$, $f_{\mathbf{t}}^{sp}$, for sketch, photo and text), and a shared modality-agnostic component (f^{ag}). The idea is that modality-specific will hold information specific to each modality (e.g., drawing style for sketch, texture for photo, and grammatical knowledge for text). It follows that filtering away modality-specific parts from *each* of the three modalities gives a shared modality-agnostic part that carries shared abstract semantic across *all* three modalities, (as shown in Fig. 1(b)). How optionality is supported in such a disentangled space then becomes trivial (Fig. 1(c),(d)). To achieve optionality across tasks, we simply use modality-agnostic information as the joint embedding to perform discriminative tasks (e.g., cross-modal retrieval), and for cross-modal generative tasks (e.g., captioning), we just combine modality-agnostic information (from source) with modality-specific (from target) to generate the target modality. Optionality across modality is a little harder, where we make use of a cross-attention [51] mechanism to capture the synergy across the modality-agnostic components.

Benefiting from our optionality-enabled embedding, we can perform a multi-facet of tasks without any task-specific modifications: (i) Fig. 1 (c) show cross-modal discriminative tasks such as sketch-based image retrieval (SBIR) using ($f_{\mathbf{s}}^{ag} \leftrightarrow f_{\mathbf{p}}^{ag}$), text-based image retrieval (TBIR) using ($f_{\mathbf{t}}^{ag} \leftrightarrow f_{\mathbf{p}}^{ag}$), or sketch+text based image retrieval (STBIR) using ($f_{\mathbf{s}}^{ag} + f_{\mathbf{t}}^{ag} \leftrightarrow f_{\mathbf{p}}^{ag}$). (ii) Fig. 1 (d) show cross-modal generative tasks such as image captioning (photo branch) using $f_{\mathbf{p}}^{ag} + f_{\mathbf{t}}^{sp} \rightarrow f_{\mathbf{t}}$ to generate textual descriptions $f_{\mathbf{t}}$. Similarly, for sketch captioning (sketch branch) we use $f_{\mathbf{s}}^{ag} + f_{\mathbf{t}}^{sp} \rightarrow f_{\mathbf{t}}$. (iii) Last but not least, to demonstrate what the expressiveness of human sketch can bring to scene understanding, we introduce a novel task of subjective captioning where we guide image captioning using sketch as a signal (subjective branch) as $f_{\mathbf{p}}^{ag} + f_{\mathbf{s}}^{ag} \rightarrow f_{\mathbf{t}}$.

In summary, our contributions are: (i) We extend multi-modal scene understanding to include human scene-sketches, thereby completing a trilogy of scene representation from three diverse and complementary modalities. (ii) We provide optionality to end-users by learning a flexible joint embedding that supports: optionality across modalities and optionality across tasks. (iii) Using computationally efficient techniques like information bottleneck, conditionally invertible neural networks, and modified cross-attention mechanism, we model this flexible joint embedding. (iv) Once learned, our embedding accommodates a multi-facet of scene-related tasks like retrieval, captioning.

2. Related Works

Sketch for Visual Understanding: Hand-drawn sketches enriched with human visual perception cues have facilitated several downstream visual understanding tasks. Apart from the widely explored SBIR [11, 23], sketch has shown potential on object localisation [19], segmentation [74], image/video synthesis [49], representation learning [81], 3D shape retrieval/modelling [21], medical image analysis [48, 102], etc. [107]. Sketches are also useful in the creative industry like artistic image editing [110] and animation [105]. Unlike photos that are passively captured by a camera, sketches are drawn by humans that actively stimulate intelligence with pictionary-style drawing games [9]. While text has been widely used for human expression, in this paper, we show freehand sketches can provide complimentary or symbiotic information for visual understanding.

Sketch-Based Image Retrieval (SBIR): SBIR retrieves a paired photo given a query sketch. Sketches offer visual description that commences the avenues of *category-level* [28, 80, 111] or fine-grained *instance-level* (FG-SBIR) [7, 10, 13] retrieval. SBIR typically employs deep triplet-ranking based siamese networks to learn a joint embedding space [112]. Contemporary research emerged towards zero-shot SBIR [28, 82], cross-domain translation [68], on-the-fly retrieval [13], semi-supervised [7], self-supervised [8], meta-learning [12] etc. As research on object-level SBIR matured, focus shifted towards the more practical scene-level SBIR [78] with GCN [58], and optimal transport [18]. The onset of scene sketch datasets [20, 35, 120] revealed further insights into implicit human-sketching strategies [20].

Text-Based Image Retrieval (TBIR): Learning image-text joint embedding space with ranking loss [32, 44, 72] received considerable attention. Further improvements used mining hardest negative pairs for triplet loss [34], cross-modal adaptive message passing [103], probabilistic one-to-many representations [22] etc. Despite text lacking visual cues, million-scale paired image-text datasets have made TBIR competitive due to power scaling laws [64]. This inspired large-scale methods like Oscar [53], and CLIP [75]. In this paper, we augment TBIR with sketches to provide the creativity and freedom of expression intrinsic to sketches.

Multi-Modality in Computer Vision: Multi-modal learning (MML) aims at developing models that can extract, interpret, and reason on information from various modalities characterised by different statistical properties such as text, sketch, or text+sketch. Contemporary research studied MML in vision via image and text [41], image to scene graph [37], etc. [108]. MML faces challenges like cross-modal alignment [43], or efficiency over data [95] and compute [45]. It is useful when data in one modality is inaccessible [3] for privacy or logistic reasons (e.g., hospital), but abundantly available in other modalities (photos in MS-COCO [56]). Often, some modalities are preferred over others for human-machine communication, like some concepts are easier to express in texts [61], while others prefer sketches [55] or both [78] (Fig. 1). In this paper, we learn cross-modal representation [109] that works using either one modality (text/sketch) or both.

Disentangled Representation for Multi-modality: Disentangling modality-agnostic from modality-specific residual factors is important for MML [42, 97]. Modality-agnostic information is useful for cross-modal transfer like semantics-based retrieval and pattern recognition [42] but holds no meaning for tasks specific to one modality like image-style or speaker information [94]. Disentanglement was explored where factors of variation are either known (e.g., facial poses [93]) and individually supervised [77], partially known [84], or unknown (e.g., drawing style [84]) and learned unsupervised using isotropic Gaussian prior [79] or information-theoric regularisation [17]. Our method aligns with the unknown setup where factors particular to sketch, text, and image are discovered unsupervised.

Image Captioning: This has emerged from predicting syntactically correct descriptions [92, 117] to tackling data scarcity [1, 50], and addressing user requirements [70, 71]. Predicted captions evolved from being factual in a neutral tone to (i) controllable using textual verbs [16], part-of-speech tag [27], or mouse trace [65, 73], and (ii) personalised captioning [87, 116] that learns user's active vocabulary, and writing style. Our method can (i) generate factual captions from images/sketches and (ii) extend controllable captioning paradigm by injecting saliency via sketch.

3. Pilot Study

3.1. Sketch vs. Text for Retrieval

Text can convey colour information, or object categories, but is cumbersome to describe fine-grained details, multiple objects, or complex shapes [89][1]. While sketch can depict complex shapes, multiple objects, and spatial alignment [20], not all objects are easy to draw ('donkey' vs. 'horse'). Fig. 2 shows this trade-off between sketch vs. text for image retrieval. We find an optimal fusion between sketch and text to derive best of both modalities along with the ability to optionally use only sketch, only text, or both.

[1]Example: Cross strap stud and buckle detail blonde leather upper leather insole chunky wooden sole 9 cm heel.

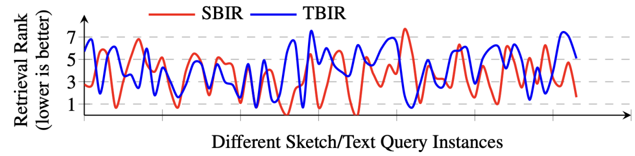


Figure 2. We compare SBIR [112] vs. TBIR [75] on FS-COCO [20] where retrieval rank is plotted in *log-scale* (see Supplemental for more details). While sketch is a better query for some instances (lower retrieval rank), for others text is better.

Table 1. Comparing alternative guiding signal like POS (part-of-speech) [27], Mouse Trace [65], and Freehand Sketches [20].

Signal		B-1	B-4	M	R	C	S
POS [27]	w/o	73.2	31.1	24.5	52.8	100.1	17.9
	w/	73.9	31.6	25.5	53.2	104.5	18.8
Δ		**0.7**	**0.5**	**1.0**	**0.4**	**4.4**	**0.9**
Trace [65]	w/o	32.2	8.1	–	31.7	29.3	25.7
	w/	52.2	24.6	–	48.3	106.5	36.5
Δ		**20**	**16.5**	**–**	**16.6**	**77.2**	**10.8**
Sketch	w/o	74.7	31.8	24.7	53.8	105.5	18.8
	w/	81.3	42.7	30.1	61.6	121.6	23.5
Δ		**6.6**	**10.9**	**5.4**	**7.8**	**16.1**	**4.7**

3.2. Subjectivity for Captioning

Unlike traditional image captioning [63, 101] that generates factual captions in neutral tone, subjective captioning adapts the predicted captions using a guiding signal that specifies priorities on what should be described [92]. The signal is injected via feature concatenation [27], or cross-attention mechanism [65]. Applications of subjective captioning include medical report generation using disease tags to generate real style reports [57], art descriptions [6], and assistive technologies for the visually impaired [38, 104]. In this paper, we advocate for sketch as a guiding signal to depict salient objects and express artistic interpretations [39]. We compare the performance (see supplementary for details) using guiding signals like POS (parts-of-speech) [27], mouse trace [65], or freehand sketches [20]. Following [65], we inject the guiding signal into the image captioning pipeline via cross-attention mechanism. As evident from Table 1, while sketch is competitive with mouse traces, it is a better signal than POS. However, unlike mouse trace, sketch can depict artistic interpretation [6] making it a more flexible and robust guiding signal than POS or mouse trace.

4. Proposed Methodology

4.1. Preliminaries

Baseline for Fine-Grained Retrieval: Given a query-photo pair $(\mathbf{q}, \mathbf{p})$, existing methods encode [8, 52, 55, 58,

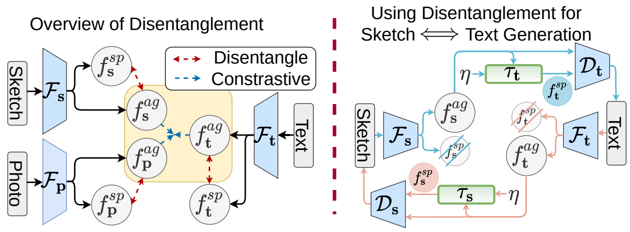


Figure 3. (Left): We disentangle modality-agnostic and modality-specific components from sketch, text, and photo. The modality-agnostic components are aligned using contrastive loss for cross-modal transfer. (Right): Modality-agnostic sketch ($f_{\mathbf{s}}^{ag}$) is used across modality to generate modality-specific text ($f_{\mathbf{t}}^{sp}$) using text-specific $\tau_{\mathbf{t}}$. Combining $f_{\mathbf{s}}^{ag}$ and $f_{\mathbf{t}}^{ag}$, we generate text from sketch.

112] the query $\mathbf{q} = \{\mathbf{s}, \mathbf{t}\}$ comprising sketch ($\mathbf{s}$) / text ($\mathbf{t}$) and photo ($\mathbf{p}$) as $f_{\mathbf{q}} = \mathcal{F}_{\mathbf{q}}(\mathbf{q}) \in \mathbb{R}^D$, and $f_{\mathbf{p}} = \mathcal{F}_{\mathbf{p}}(\mathbf{p}) \in \mathbb{R}^D$ respectively. The network is trained via triplet loss with margin parameter $\mu > 0$ such that the cosine distance $\delta(\cdot)$ of query anchor $\mathbf{q}$ from a negative photo ($\mathbf{p}^-$) should increase while that from the positive photo ($\mathbf{p}^+$) should decrease as, $\mathcal{L}_{trip} = \max\{0, \mu + \delta(f_{\mathbf{q}}, f_{\mathbf{p}^+}) - \delta(f_{\mathbf{q}}, f_{\mathbf{p}^-})\}$.

Baseline for Image Captioning: Image captioning consists of an image encoder [60, 106], $f_{\mathbf{p}} = \mathcal{F}_{\mathbf{p}}(\mathbf{p})$ followed by an autoregressive textual decoder ($\mathcal{F}_C$). Given the textual description comprises a sequence of words $\mathbf{t} = \{w_1, \dots, w_K\}$, we maximise the likelihood of a predicted word ($\hat{w}_k$) at each step (k), conditioned on $f_{\mathbf{p}}$ as, $\mathcal{L}_C = -\sum_{k=1}^{K} \log[\mathcal{F}_C(\hat{w}_k = w_k | f_{\mathbf{p}}, w_1, \dots, w_{k-1})]$

4.2. Overview

We aim to disentangle the feature representations from sketch, text, and photo modalities into a *modality-agnostic* and *modality-specific* component. While the modality-*agnostic* component holds semantic information to support cross-modal transfer, the modality-*specific* one holds information necessary during self-reconstruction; however, it lacks meaning in other modalities (e.g., grammatical knowledge in text). Achieving feature disentanglement across scene sketches, texts, and photos enables a multitude of downstream tasks like (i) *SBIR* – modality-agnostic sketch and photo features, (ii) *TBIR* – modality-agnostic text and photo, (iii) *Sketch+Text-Based Image Retrieval* – modality-agnostic sketch, text, and photo, (iv) *Image Captioning* – using the modality-agnostic photo to compute modality-specific text features, (v) *Sketch Captioning* – modality-agnostic sketch to compute modality-specific text, and (vi) *Subjective Captioning* – using modality-agnostic photo and sketch, to compute modality-specific text.

4.3. Disentangling Modality Agnostic and Specific

While our disentangling method can be generalised to any number of modalities, for simplicity, we first show for $M = 2$ modalities and later extend to $M \geq 3$. Consider a simple bimodal setup of sketch ($\mathbf{s} \in \mathbb{R}^{H \times W \times 3}$) and text ($\mathbf{t} \in \mathbb{R}^{N \times E}$). Our goal is to split the feature representation $f_{\mathbf{s}} = \mathcal{F}_{\mathbf{s}}(\mathbf{s}) \in \mathbb{R}^{512}$ and $f_{\mathbf{t}} = \mathcal{F}_{\mathbf{t}}(\mathbf{t}) \in \mathbb{R}^{512}$ into a modality-**ag**nostic and a modality-**sp**ecific component as $f_{\mathbf{s}} = [f_{\mathbf{s}}^{ag}, f_{\mathbf{s}}^{sp}]$, and $f_{\mathbf{t}} = [f_{\mathbf{t}}^{ag}, f_{\mathbf{t}}^{sp}]$ respectively, where $f^{ag} \in \mathbb{R}^{480}$ and $f^{sp} \in \mathbb{R}^{32}$. Existing methods [84, 91] disentangle feature representations via (i) self reconstruction as $\hat{\mathbf{s}} = \mathcal{D}_{\mathbf{s}}([f_{\mathbf{s}}^{ag}, f_{\mathbf{s}}^{sp}])$ and $\hat{\mathbf{t}} = \mathcal{D}_{\mathbf{t}}([f_{\mathbf{t}}^{ag}, f_{\mathbf{t}}^{sp}])$ coupled with (ii) cross-modal translation $\hat{\mathbf{s}} = \mathcal{D}_{\mathbf{s}}([f_{\mathbf{t}}^{ag}, f_{\mathbf{s}}^{sp}])$ and $\hat{\mathbf{t}} = \mathcal{D}_{\mathbf{t}}([f_{\mathbf{s}}^{ag}, f_{\mathbf{t}}^{sp}])$. However, using cross-modal translation with latent feature exchange across modalities is a cumbersome process that explodes with $\mathbb{P}_2^M$ permutations for M modalities, e.g., $M = 3$ has $\mathbb{P}_2^3 = 6$ cross-modal translations. Adding multiple cross-modal translation losses makes optimisation difficult and computationally expensive. We break this compute barrier with linear ($\mathcal{O}(M)$) complexity using an *information bottleneck reinterpretation* of modality-agnostic and modality-specific disentanglement. In particular, we *maximise* the *mutual information* $\mathcal{I}(f_{\mathbf{s}}^{ag}, f_{\mathbf{t}}^{ag})$ amongst modality-agnostic components, while *minimising* the *same* between modality-agnostic and modality-specific components $\mathcal{I}(f_{\mathbf{s}}^{ag}, f_{\mathbf{s}}^{sp})$, and $\mathcal{I}(f_{\mathbf{t}}^{ag}, f_{\mathbf{t}}^{sp})$, where $\mathcal{I}\ (\cdot,\cdot)$ denotes mutual information between two entities. Hence, unlike the previous $\mathbb{P}_2^M$ permutations, Eq. (1) has one agnostic $\mathcal{I}(f_{\mathbf{s}}^{ag}, f_{\mathbf{t}}^{ag})$, and M specific $\mathcal{I}(f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp})$ losses. Formally, using a Langrange multiplier hyperparameter β we have our loss objective as,

$$\mathcal{L}_{\mathcal{I}} = -\overbrace{\mathcal{I}(f_{\mathbf{s}}^{ag}, f_{\mathbf{t}}^{ag})}^{agnostic} + \beta \overbrace{\sum_{\mathbf{k} \in \{\mathbf{s}, \mathbf{t}\}} \mathcal{I}(f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp})}^{specific} \quad (1)$$

Minimise $\mathcal{I}(f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp})$: We minimise the mutual information between modality-agnostic and modality-specific components using a conditional invertible [] neural network $\tau_{\mathbf{k}}$. Unlike typical unidirectional neural networks $\mathcal{F} : x \rightarrow y$, a conditional invertible neural network employs a sequence of bijective mapping operations like activation normalization (ActNorm) [46], Conditional Affine Coupling [29], and shuffling [46] to obtain $\tau_{\mathbf{k}} : x \leftrightarrow y$. During the *forward pass* (inference), we sample $\eta \in \mathbb{R}^{32}$ from a uniform prior distribution $\mathbb{p}(\eta)$ to predict the modality-specific $f_{\mathbf{k}}^{sp} \in \mathbb{R}^{32}$ by conditioning on $f_{\mathbf{k}}^{ag}$ as, $f_{\mathbf{k}}^{sp} =$

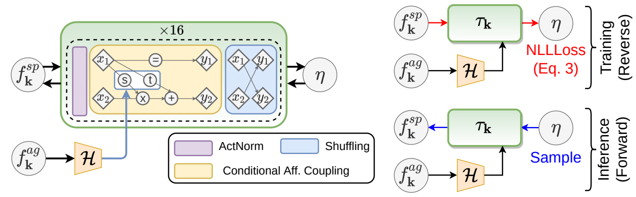


Figure 4. Unlike typical neural networks that are unidirectional, data in conditional invertible neural networks ($\tau_{\mathbf{k}}$) can flow either (i) from modality-specific $f_{\mathbf{k}}^{sp}$ to a uniform distribution η by conditioning on modality-agnostic $f_{\mathbf{k}}^{ag}$ **during training**, or (ii) from a sampled η in uniform distribution to the modality-specific $f_{\mathbf{k}}^{sp}$ by conditioning on modality-agnostic $f_{\mathbf{k}}^{ag}$ **during inference**. The conditioning vector $f_{\mathbf{k}}^{ag}$ is injected into the conditional affinity coupling layers [29] of $\tau_{\mathbf{k}}$ using any arbitrary network $\mathcal{H}$.

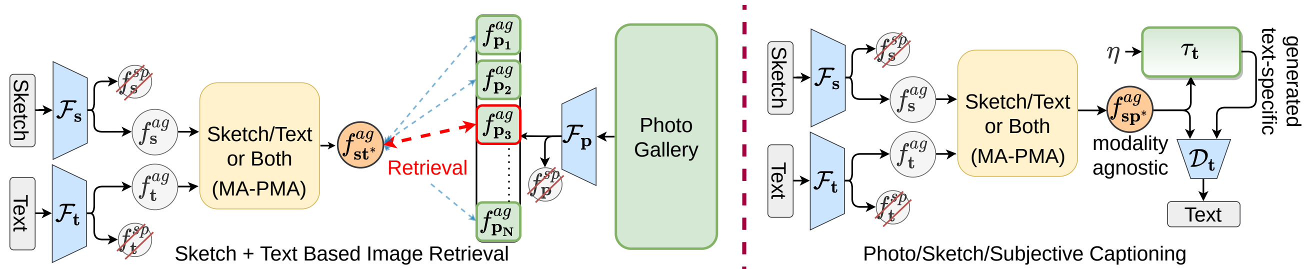


Figure 5. (Left): The modality-agnostic from sketch, or text, or both are used to retrieve from a gallery of photos. This enables a multitude of retrieval tasks like SBIR, TBIR, and STBIR. (Right): The modality-agnostic from photo, or sketch, or both are used to generate the text-specific component. Combining the modality-agnostic and inferred text-specific (via $\tau_{\mathbf{t}}$) enables image, or sketch, or subjective captioning.

$\tau_{\mathbf{k}}(\eta \mid f^{ag}_{\mathbf{k}})$. In other words, during inference, we predict the modality-specific component of target from the modality-agnostic one of input using $\tau_{\mathbf{k}}$. The target modality is then generated by combining the input-agnostic and target-specific components. The conditioning modality-agnostic vector $f^{ag}_{\mathbf{k}}$ is injected into the intermediate conditional affine coupling layers $\mathcal{C}: x \leftrightarrow y$ as: $[x_1, x_2] = \texttt{split}(x)$, and $y = \texttt{concat}[x_1, s_\theta([x_1; h]) \odot x_2 + t_\theta([x_1; h])]$, where, $h = \mathcal{H}(f^{ag}_{\mathbf{k}})$. A simple feed-forward neural network implements s_θ, t_θ, and $\mathcal{H}$. We learn $\tau_{\mathbf{k}}$ in the *reverse pass* (training) via negative log-likelihood (NLL Loss in Fig. 4) of $\tau^{-1}_{\mathbf{k}}(f^{sp}_{\mathbf{k}} \mid f^{ag}_{\mathbf{k}})$ to predict a uniform distribution $\mathbb{p}(\eta)$,

$$\mathbb{p}(\eta) = \mathbb{p}(\tau^{-1}_{\mathbf{k}}(f^{sp}_{\mathbf{k}} \mid f^{ag}_{\mathbf{k}}))\ |\text{det}J_{\tau^{-1}_{\mathbf{k}}}(f^{sp}_{\mathbf{k}} \mid f^{ag}_{\mathbf{k}})| \qquad (2)$$

We show how learning $\tau_{\mathbf{k}}$ in Eq. (2) minimises $\mathcal{I}(f^{ag}_{\mathbf{k}}, f^{sp}_{\mathbf{k}})$. $\mathcal{I}(f^{ag}_{\mathbf{k}}, f^{sp}_{\mathbf{k}}) = \int_{f^{sp}_{\mathbf{k}}} \mathbb{p}(f^{sp}_{\mathbf{k}} | f^{ag}_{\mathbf{k}}) \log \mathbb{p}(f^{sp}_{\mathbf{k}} | f^{ag}_{\mathbf{k}}) / \mathbb{p}(f^{sp}_{\mathbf{k}})$. Approximating modality-specific prior $\mathbb{p}(f^{sp}_{\mathbf{k}})$ with variational distribution $\mathbb{q}(f^{sp}_{\mathbf{k}})$ gives the upper-bound, minimising which reduces the KL-divergence between $\mathbb{p}(f^{sp}_{\mathbf{k}} | f^{ag}_{\mathbf{k}})$ and $\mathbb{q}(f^{sp}_{\mathbf{k}})$ i.e., it encourages the disentanglement $\mathbb{p}(f^{ag}_{\mathbf{k}}, f^{sp}_{\mathbf{k}}) \approx \mathbb{p}(f^{ag}_{\mathbf{k}}) \cdot \mathbb{p}(f^{sp}_{\mathbf{k}})$. The prior $\mathbb{q}(f^{sp}_{\mathbf{k}})$ is solved using $\tau_{\mathbf{k}}$ to enforce disentanglement between modality-agnostic and modality-specific components, like that in Eq. (2), as the sum of *negative-loglikelihood* (NLL-Loss in Fig. 4) and *log-determinant* (see supplementary for proof),

$$\begin{aligned}\mathcal{L}_{\tau_{\mathbf{k}}} = &-\mathbb{E}_{f^{sp}_{\mathbf{k}}}\{\log \mathbb{q}(\tau^{-1}_{\mathbf{k}}(f^{sp}_{\mathbf{k}} \mid f^{ag}_{\mathbf{k}})) \\ &+ \log|\text{det}J_{\tau^{-1}_{\mathbf{k}}}(f^{sp}_{\mathbf{k}} \mid f^{ag}_{\mathbf{k}})|\}\end{aligned} \qquad (3)$$

Maximise $\mathcal{I}(f^{ag}_{\mathbf{s}}, f^{ag}_{\mathbf{t}})$: Here we show how minimising a constrastive based retrieval loss [98] between the modality-agnostic components of sketch and text will maximise their mutual information. We define contrastive loss matching modality-agnostic components of sketch and text as,

$$\mathcal{L}^{\mathbf{s},\mathbf{t}}_{cl} = -\mathbb{E}_{f^{sp}_{\mathbf{s}}}\left[\log \frac{\omega(f^{ag}_{\mathbf{s}}, f^{ag}_{\mathbf{t}^+})}{\omega(f^{ag}_{\mathbf{s}}, f^{ag}_{\mathbf{t}^+}) + \sum^{N-1}_{f^{ag}_{\mathbf{t}^-}} \omega(f^{ag}_{\mathbf{s}}, f^{ag}_{\mathbf{t}^-})}\right] \qquad (4)$$

where, $\omega = \exp(x^T\ \mathbf{W}\ y)$. For each modality-agnostic $f^{ag}_{\mathbf{s}}$ we sample a positive $f^{ag}_{\mathbf{t}^+}$ and $(N-1)$ negative $f^{ag}_{\mathbf{t}^-}$ pairs. The contrastive loss in Eq. (4) is expressed as mutual information between $f^{ag}_{\mathbf{s}}$ and $f^{ag}_{\mathbf{t}}$ as, $\mathcal{L}^{\mathbf{s},\mathbf{t}}_{cl} \geq -\mathcal{I}(f^{ag}_{\mathbf{s}}, f^{ag}_{\mathbf{t}}) + \log(N)$. Hence, to maximise the mutual information between modality-agnostic $f^{ag}_{\mathbf{s}}$ and $f^{ag}_{\mathbf{t}}$, we can maximise the tractable lower bound $\log(N) - \mathcal{L}^{\mathbf{s},\mathbf{t}}_{cl}$.

Total Loss for Bimodal Setup: The resulting loss ($\mathcal{L}_{tot}$) for bimodal (sketch and text) setup comprise three loss objectives (i) *self reconstruction* loss $\mathcal{L}_{rec}$, (ii) *contrastive loss* between two modality-agnostic terms $\mathcal{L}^{\mathbf{s},\mathbf{t}}_{cl}$, and (iii) *disentanglement* between modality-agnostic and modality-specific components in each modality ($\mathbf{k}$) ($\mathcal{L}_{\tau_{\mathbf{k}}}$), as

$$\begin{aligned}\mathcal{L}_{rec} &= ||\mathbf{s} - \mathcal{D}_{\mathbf{s}}(\mathcal{F}_{\mathbf{s}}(\mathbf{s}))||_2 + ||\mathbf{t} - \mathcal{D}_{\mathbf{t}}(\mathcal{F}_{\mathbf{t}}(\mathbf{t}))||_2 \\ \mathcal{L}_{tot} &= \mathcal{L}_{rec} + \mathcal{L}^{\mathbf{s},\mathbf{t}}_{cl} + \beta[\mathcal{L}_{\tau_{\mathbf{s}}} + \mathcal{L}_{\tau_{\mathbf{t}}}]\end{aligned} \qquad (5)$$

Extending to Three/More Modalities: Here we extend our bimodal setup in Sec. 4.3 to three or more modalities. (i) We compute the self-reconstruction loss for three modalities as $\mathcal{L}_{rec} = \sum_{\mathbf{k}\in\{\mathbf{s},\mathbf{t},\mathbf{p}\}} ||\mathbf{k} - \mathcal{D}_{\mathbf{k}}(\mathcal{F}_{\mathbf{k}}(\mathbf{k}))||_2$. (ii) we minimise the mutual information between modality-agnostic and modality-specific components for sketch, text, and photo as, $\mathcal{L}_\tau = \mathcal{L}_{\tau_{\mathbf{s}}} + \mathcal{L}_{\tau_{\mathbf{t}}} + \mathcal{L}_{\tau_{\mathbf{p}}}$. (iii) However, our contrastive loss term $\mathcal{L}_{cl}$ that maximises the mutual information among modality-agnostic components can only compare two modalities. We can extend this naively to a three-modality setup as $\mathcal{L}^{tot}_{cl} = \mathcal{L}^{\mathbf{s},\mathbf{t}}_{cl} + \mathcal{L}^{\mathbf{s},\mathbf{p}}_{cl} + \mathcal{L}^{\mathbf{t},\mathbf{p}}_{cl}$.

Extending to three or more modalities, however, we notice our contrastive loss in Eq. (4) is defined for only bimodal setup ($\mathcal{L}^{\mathbf{s},\mathbf{t}}_{cl}$, or $\mathcal{L}^{\mathbf{s},\mathbf{p}}_{cl}$, or $\mathcal{L}^{\mathbf{t},\mathbf{p}}_{cl}$). For example, given three modalities $\mathcal{S}_M = \{m_1, m_2, m_3\}$, comparing only (m_1, m_2) ignores m_3. This highlights a key limitation: it fails when we have a query in both (m_1, m_3) to retrieve m_2 (e.g., sketch+text for image retrieval). Now the research question boils down to – how can we model a function $\mathcal{G}(\cdot)$ such that it can model either m_1, or m_3, or both (m_1, m_3) to retrieve m_2. To design $\mathcal{G}$, using naive addition as $\mathcal{G}(m_1, m_3) = m_1 + m_3$ does not handle overlapping or conflicting information[2] in m_1 and m_3 [59]. While, concatenation $\mathcal{G}(m_1, m_3) = \texttt{concat}[m_1, m_3]$ computes interaction between (m_1, m_3), it forces to provide both m_1

[2]When signals (m_1, m_3) are similar or complementary $\mathcal{G}$ should strengthen decision; when signals conflict $\mathcal{G}$ should filter unreliable ones.

and m_3 during inference; thereby failing to model either m_1, or m_3, or both (m_1, m_3).

4.4. Modelling Optional Sketch or Text

We propose a simple approach to design $\mathcal{G}$ that optionally models either m_1, or m_3, or both (m_1, m_3), and handles overlapping or conflicting information. Our proposed $\mathcal{G}$ comprises a multihead cross-attention module $\texttt{MH}(\cdot)$ followed by an attention-based pooling $\texttt{PMA}(\cdot)$ as, $f_M = \texttt{PMA}(H_M)$; where $H_M = \texttt{MH}(\mathcal{S}_M)$, and $\mathcal{S}_M = \{m_1, m_3\}$.

Our $\texttt{MH}(\cdot)$ is order-invariant and independent of the number (M) of input modalities defined as $\texttt{MH}(X) = \sigma(XX^T)X$; where σ is scaled-softmax, X^T is transpose of X, and $X \in \mathbb{R}^{M\times 480}$ is a list of modality-agnostic components m_1, or m_3 with $\mathbb{R}^{1\times 480}$, or $(m_1, m_3) \in \mathbb{R}^{2\times 480}$ in query. The cross-attention in $\texttt{MH}(\cdot)$ interacts across query modalities to compute mutually agreeing information between (m_1, m_3) as, $H_M \in \mathbb{R}^{2\times 480}$. Next, we use an order-invariant attention-based pooling $\texttt{PMA} : \mathbb{R}^{2\times 480} \rightarrow \mathbb{R}^{1\times 480}$ with a learned seed vector $\mathcal{P} \in \mathbb{R}^{1\times 480}$ to aggregate mutually agreeing H_M as, $f_M = \texttt{PMA}(H_M) = \sigma(\mathcal{P}H_M^T)H_M$. Hence, using our proposed fusion module $\mathcal{G}$, we adapt our contrastive loss defined for only a pair of modality-agnostic components in Eq. (4) as $\mathcal{L}_{cl}^{tot} = \mathcal{L}_{cl}^{\mathbf{s},\mathbf{t}} + \mathcal{L}_{cl}^{\mathbf{s},\mathbf{p}} + \mathcal{L}_{cl}^{\mathbf{t},\mathbf{p}}$ to jointly model sketch–text–photo (or more) modality-agnostic as: $\mathcal{L}_{cls}^{tot} = \mathcal{L}_{cl}(\mathcal{G}(f_{\mathbf{s}}^{ag}, f_{\mathbf{t}}^{ag}), f_{\mathbf{p}}^{ag}) + \mathcal{L}_{cl}(\mathcal{G}(f_{\mathbf{s}}^{ag}, f_{\mathbf{p}}^{ag}), f_{\mathbf{t}}^{ag}) + \mathcal{L}_{cl}(\mathcal{G}(f_{\mathbf{p}}^{ag}, f_{\mathbf{t}}^{ag}), f_{\mathbf{s}}^{ag})$. For a generalised solution involving more than three modalities ($M > 3$), see supplementary.

Inference Data Flow: We describe the inference data flow in Fig. 5. For retrieval tasks, we first compute the modality-agnostic component of query sketch and text $(f_{\mathbf{s}}^{ag}, f_{\mathbf{t}}^{ag})$, and a gallery of photos $\{f_{\mathbf{p_1}}^{ag}, f_{\mathbf{p_2}}^{ag}, \dots, f_{\mathbf{p_N}}^{ag}\}$. Next, a combined representation for either only sketch $(f_{\mathbf{s}}^{ag})$, or only text $(f_{\mathbf{t}}^{ag})$, or both $(f_{\mathbf{s}}^{ag}, f_{\mathbf{t}}^{ag})$ is computed using multihead cross attention $\texttt{MH}(\cdot)$ followed by attention-based pooling $\texttt{PMA}(\cdot)$ defined in Sec. 4.4 to get $f_{\mathbf{st}*}^{ag}$. Finally, we find the minimum distance between the combined $f_{\mathbf{st}*}^{ag}$ and modality-agnostic component of photo $f_{\mathbf{p}_i}^{ag}$ as $\omega(f_{\mathbf{st}*}^{ag}, f_{\mathbf{p_i}}^{ag})$ defined in Eq. (4). For captioning, we additionally use the text-specific conditional invertible neural network $\tau_{\mathbf{t}}$ to generate the target modality-specific text (e.g., grammatical structure etc.) from input modality-agnostic comprising of only photo $(f_{\mathbf{p}}^{ag})$ for image captioning, only sketch $(f_{\mathbf{s}}^{ag})$ for sketch captioning, or both photo and sketch $(f_{\mathbf{p}}^{ag}, f_{\mathbf{s}}^{ag})$ for subjective captioning (i.e., generate image captions by conditioning on the input sketch).

5. Experiments

Datasets: We use two scene sketch datasets with fine-grained alignment among sketch, text, and photo: (i) SketchyCOCO [35] contains $14,081$ scene sketch-photo pairs. The photos are taken from MS-COCO [56] comprising $164K$ photos with paired texts. However, most sketches in SketchyCOCO [35] contain less than one foreground instance. Following [58], we filter SketchyCOCO with one foreground instance to get $1015/210$ train/test scene sketches. (ii) Unlike SketchyCOCO [35], where the scene sketches are synthetically generated, FS-COCO [20] includes $7000/3000$ train/test human-drawn scene sketches with a paired textual description of sketches.

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 for image captioning using $82,783$ photo-text pairs (excluding the photos common in SketchyCOCO and FS-COCO) for 15 epochs. Next, we fine-tune on either SketchyCOCO [35], or FS-COCO [20] for 200 epochs using Adam optimiser with learning rate $1e-4$, and batch size 64. Our photo $(\mathcal{F}_{\mathbf{p}})$ and sketch $(\mathcal{F}_{\mathbf{s}})$ encoders use ImageNet pretrained VGG-16 [88]. For simplicity, we encode text using a bidirectional GRU unit with 512 hidden units. Our text decoder [44] is a single-layer autoregressive LSTM decoder that predicts a probability distribution over a fixed vocabulary $(10,010$ words) at every time step. For the image/sketch decoder, we use two separate GAN [114] networks that synthesise sketch/image of size 64×64, respectively. For brevity, we avoid realistic sketch/image generation due to the challenging scene complexity [20]. Hence we do not use a discriminator module for high-quality, sharp reconstruction [115]. Finally, our conditionally invertible neural network comprises 16 alternating affine coupling [30], activation normalisation [46], and switch layers [30].

Evaluation Metric: In line with FG-SBIR research, we use Acc.@q [83] defined as the percentage of sketches having a true matched photo in the top-q list. For sketch/image/subjective captioning, we use standard metrics BELU (B) 1-4 [69], CIDEr (C) [99], ROUGE (R) [54], METEOR (M) [26], and SPICE [2]. Following [101], we generate 100 candidate captions and employ consensus reranking using CIDEr to select the best candidate caption.

Competitors: We compare against (i) existing state-of-the-art methods that align two modalities (S2): For SBIR, **Triplet-SN** [112] employs Sketch-A-Net [113] backbone trained using triplet loss. **HOLEF** [90] adds spatial attention with a higher-order ranking loss. **SketchyS** [120] replaces Sketch-A-Net in *Triplet-SN* with VGG-16 [88] and an auxiliary category-level cross-entropy. **SceneS** [58] uses GCN [47] to model scene sketch layout information. For TBIR, **CLIP** [75] is trained with text using transformer [86] and photo using vision transformer [31] on 400 million text-photo pairs. **CLIP-LN** fine-tunes *CLIP* by training only layer normalisation parameters [5] with learning rate 0.00001. For image/sketch captioning, **SAT** [106] is one of the simplest but seminal works using a CNN-LSTM encoder-decoder approach similar to ours. **GMM-CVAE**

Table 2. Quantitative results combining sketch and text for image retrieval (FG-STBIR) on two scene sketch datasets [20, 35].

	Method	SketchyCOCO [35]		FSCOCO [20]	
		Acc.@1	Acc.@10	Acc.@1	Acc.@10
S3	QST [89]	38.9	87.9	25.1	54.5
	SCM [3]	38.5	87.3	24.3	54.1
B	CrossAtt [86]	39.1	88.2	25.3	54.8
	Proposed	**39.5**	**88.7**	**25.7**	**55.2**

[101] employs a conditional variational autoencoder with a Gaussian mixture model. **LNFMM** [62] is similar to ours that splits information into modality-agnostic and modality-specific components using conditional invertible neural network, **ClipCap** [66] employs CLIP [75] for image encoding followed by GPT-2 [76] for text decoding. A learned mapping module translates CLIP embeddings to GPT-2. (ii) We compare against methods that align 3 modalities (S3): For STBIR, **QST** [89] extends triplet loss in *Triplet-SN* to quadruplet loss that combines sketch and text for image retrieval. **SCM** uses element-wise addition to combine sketch and text from ResNet-18 [40] with weight sharing across sketch, text, and photo from *ResBlock4* onwards [3]. (iii) We design baselines (B): For STBIR, **CrossAtt** employs cross-attention [86] to combine sketch and text. For subjective captioning, **MulCap** combines sketch ($f_{\mathbf{s}}^{ag}$) and photo ($f_{\mathbf{p}}^{ag}$) via element-wise multiplication as in [15]. **CrossCap** optionally fuse photo, sketch, or both using cross-attention. **CatCap** use feature concatenation [73] of guiding sketch ($f_{\mathbf{s}}^{ag}$) signal with photo ($f_{\mathbf{p}}^{ag}$) to generate captions.

5.1. Combining Sketch and Text for Image Retrieval

Fig. 2 shows that for some instances, sketch is a better query, whereas text is better for others. Hence, to achieve *best of both modalities*, we examine the complimentary nature by combining sketch and text for image retrieval. Table 2 shows (i) *SCM* gives the lowest performance due to naive element-wise addition of potentially overlapping and conflicting information [59] from sketch and text. (ii) *QST* improves slightly upon *SCM* by replacing naive element-wise addition with a weighted summation (0.8 for sketch modality). (iii) *CrossAtt* outperforms all baselines by using a cross-attention between sketch and text to resolve overlapping/conflict information [59]. (iv) Our proposed method gives the highest performance due to cross-attention that model sketch-text interaction and disentanglement to drive out modality-specific information for cross-modal retrieval.

5.2. Optionally using Sketch for Image Retrieval

Our method allows drawing only easy-to-sketch scenes instead of using both sketch and text forcibly. Table 3 compares against methods that specialise on two-modalities (S2), three-modalities (S3), and our proposed baselines (B). We observe (i) training on three modalities (sketch, text, and photo) in S3 generally outperforms those trained using only sketch and photo (S2). This can be attributed to learning generalisable features in multi-modal setup [3]. (ii) *QST* in S3 outperforms *SCM* indicating quadruplet loss is a better training objective than naive element-wise addition when combining sketch, text, and photo. (iii) Performance difference between *CrossAttn* and *QST* is not as significant as in FG-STBIR (Table 2) as during inference, we only use sketch, omitting the cross-attention module. (iv) Our method outperforms S2, S3, and B even for two-modality setup thanks to disentanglement that eliminates confounding [3] modality-specific information.

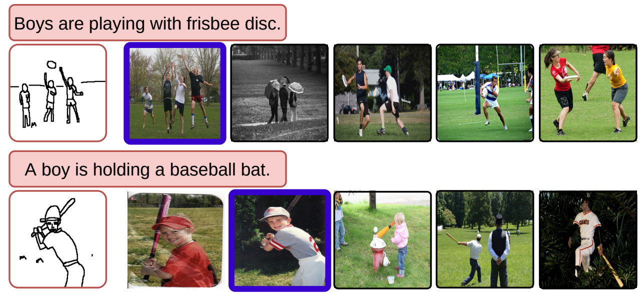


Figure 6. Qualitative results of combining sketch and text as query for image retrieval on FSCOCO [20]. See supplementary for more.

Table 3. Quantitative results using only sketch for image retrieval (FG-SBIR) on two scene sketch datasets [20, 35].

	Method	SketchyCOCO [35]		FSCOCO [20]	
		Acc.@1	Acc.@10	Acc.@1	Acc.@10
S2	Triplet-SN [112]	6.2	32.9	4.7	21.0
	HOLEF [90]	6.2	40.7	4.9	21.7
	SketchyS [120]	36.5	78.6	23.0	52.3
	SceneS [58]	31.9	86.2	–	–
S3	QST [89]	37.4	87.1	23.6	52.9
	SCM [3]	37.3	86.8	23.4	52.6
B	CrossAtt	37.9	87.4	23.7	53.5
	Proposed	**38.2**	**87.6**	**24.1**	**53.9**

Table 4. Quantitative results of fine-grained text-based image retrieval (FG-TBIR) on two scene sketch datasets [20, 35].

	Method	SketchyCOCO [35]		FSCOCO [20]	
		Acc.@1	Acc.@10	Acc.@1	Acc.@10
S2	CLIP [75]	21.0	50.9	11.5	35.3
	CLIP-LN [75]	22.1	52.3	14.8	36.6
S3	QST [89]	11.1	31.1	7.2	23.6
	SCM [3]	10.7	31.0	6.9	23.1
B	CrossAtt	20.1	51.0	12.5	35.8
	Proposed	**21.5**	**51.6**	**13.7**	**36.3**

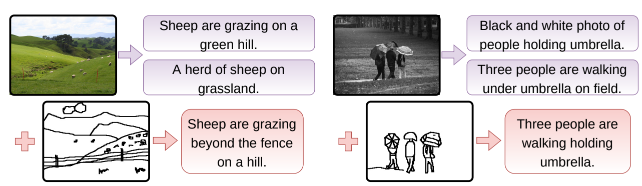


Figure 7. Qualitative results for image captioning v/s subjective captioning on FS-COCO [20]. See supplementary for more.

Table 5. Quantitative results of standard captioning metrics on MS-COCO [56] and FS-COCO [20] dataset.

		Image Captioning						Sketch Captioning						Subjective Captioning					
	Method	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S
S2	SAT [106]	71.8	25.0	23.0	–	–	–	46.2	13.7	17.1	44.9	69.4	14.5	–	–	–	–	–	–
	GMM-CVAE [101]	72.9	30.7	24.2	52.5	98.6	17.7	49.6	15.5	18.3	48.7	77.6	15.5	–	–	–	–	–	–
	AG-CVAE [101]	73.2	31.1	24.5	52.8	100.1	18.8	50.9	16.0	18.9	49.1	80.5	15.8	–	–	–	–	–	–
	LNFMM [62]	74.7	31.8	24.7	53.8	105.5	18.8	52.2	16.7	21.0	52.9	90.1	16.0	–	–	–	–	–	–
B	MulCap	74.9	33.2	25.5	54.9	106.0	19.5	53.9	17.0	21.0	53.8	97.3	16.7	78.7	38.6	28.5	59.8	110.7	21.7
	CatCap	–	–	–	–	–	–	–	–	–	–	–	–	77.6	38.0	28.3	57.7	108.0	21.2
	CrossCap	75.5	34.3	26.1	55.4	106.7	20.1	54.3	17.9	21.4	54.3	100.3	17.5	79.2	39.3	28.4	59.5	117.3	22.1
	Proposed	**76.0**	**35.9**	**26.9**	**56.9**	**107.0**	**20.9**	**56.9**	**19.3**	**21.6**	**56.6**	**106.5**	**18.9**	**81.3**	**42.7**	**30.1**	**61.6**	**121.6**	**23.5**

Table 6. Ablation study on FG-STBIR and Subjective Captioning using FSCOCO [20]. `CA` denotes cross-attention in Sec. 4.4.

$\tau_{\mathbf{k}}$	CA	$\mathcal{L}_{cl}$	Acc.@1	Acc.@10	B-1	C
✗	✗	✗	24.5	53.7	73.3	100.1
✓	✗	✗	24.9	54.0	77.9	108.5
✓	✓	✗	25.5	54.9	80.6	119.3
✓	✓	✓	**25.7**	**55.2**	**81.3**	**121.6**

5.3. Optionally using Text for Image Retrieval

While some information is best expressed by drawing, others, like colour, is best described via text. From Table 4, we observe (i) Given the same train/test split, sketches outperform text as a query modality for fine-grained image retrieval. (ii) *CLIP* and *CLIP-LN* outperforms all competitors due to superior pre-trained weights using 400 million text-image pairs. (iii) The proposed method outperforms most methods due to disentanglement that drives out modality-specific components. Although CLIP [75] outperforms the proposed method, we deliberately use a simple and easy-to-reproduce GRU/VGG-16 architectures for text/photo encoders, and train on a much smaller data [20,35] than CLIP.

5.4. Image or Sketch Captioning

In addition to disentanglement for cross-modal retrieval tasks (e.g., FG-SBIR, FG-TBIR), our conditional invertible neural network $\tau_{\mathbf{t}}$ can also generate text-specific information (Fig. 4) to support generative tasks like image/sketch captioning. We generate 100 candidate captions using (i) beam search for *SAT*, *MulCap*, *CrossCap*, *CatCap*, and (ii) sampling from prior distribution for *GMM-CVAE*, *AG-CVAE*, *LNFMM*, and our proposed method. From Table 5, we observe (i) our baselines adopting recent techniques like vision-transformer [31] outperforms (S2) – recent but complex approaches like *LNFMM*, *AG-CVAE*, and the older yet seminal work like *SAT*. (ii) Performance gap between *MulCap* and *CrossCap* is insignificant for two-modality setups (photo to text, or sketch to text) since they primarily differentiate by their multi-modal (photo and sketch) fusion strategy. (iii) In spite of using a photo/sketch encoder and text decoder similar to our simple competitor *SAT*, our proposed method performs competitively with complex methods like *LNFMM*, *AG-CVAE*, and latest approaches using vision-transformers [31], like *CrossCap*. This shows the significant contribution of (i) disentangling modality-specific and modality-agnostic components from photo/sketch, and (ii) modelling text-specific prior for generative tasks.

5.5. Sketch Based Subjective Captioning

As defined in Sec. 3.2, unlike traditional captioning frameworks that factually describe an image or sketch in a neutral tone, subjective captioning focus on drawing out a user's intentions, salient objects, and artistic interpretations [39]. Being the first method to use scene-level sketch as a guiding signal for captioning, we follow controllable captioning literature [92] to adopt three baselines (**B**) that inject the sketch conditioning signal into the captioning pipeline. From Table 5, we observe (i) *MulCap* outperforms *CatCap*, thereby supporting previous observations [15] of element-wise multiplication being more effective than concatenation. (ii) *CrossAtt* outperforms all baselines (B) and two-modality SOTAs (S2) by using a cross-attention mechanism to fuse sketch and photo by modelling sketch-photo interactions to resolve overlapping or conflicting information. Our proposed method is similar to *CrossAtt* using cross-attention (Sec. 4.4) but also enriches the modality-agnostic sketch and photo features by removing the confounding modality-specific information to offer the best performance.

5.6. Ablation

In Table 6, we evaluate the contribution of each key design choice on FG-STBIR and Subjective captioning using FS-COCO [20]. (i) Replacing cross-attention in Sec. 4.4 with quadruplet loss [89] leads to a performance drop by $0.6/0.9/2.7/10.8$ in Acc.@1/Acc.@10/B-1/C metrics respectively to show the importance of modelling the interaction between sketch and text. (ii) Replacing contrastive loss-based query-photo score in Eq. (4) with a simple triplet loss leads to a performance drop by $0.2/0.3/0.7/2.3$ due to the inability of unimodal $L2$-based triplet loss to model highly complex scene information [98]. (iii) Finally, removing the conditional invertible neural networks ($\tau_{\mathbf{k}}$) drops retrieval and captioning by $0.4/0.3/4.6/8.4$ due to percolation of the confounding modality-specific information in cross-modal tasks [3] and the inability to generate text-specific information from photo and sketch respectively.

6. Conclusion

We have studied for the first time the trilogy 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 support several downstream tasks like fine-grained sketch-based image retrieval, fine-grained sketch and text based image retrieval, sketch captioning, and subjective captioning, among others. Future research can explore challenging downstream tasks such as scene-level sketch-based image generation, sketch and text based image generation, and text-based sketch generation tasks.

A. Details for Subjective Captioning

We provide additional details of our pilot study in Sec. 3.2 that compare the performance of subjective captioning when using part-of-speech (POS) [27], mouse trace [65] or sketch as a guiding signal into the image captioning pipeline. Instead of choosing a common baseline to compare subjective captioning when using POS, mouse trace, and sketches, we measure the relative performance over the standard baselines used in recent literature to study the contribution of every guiding signal. (i) For POS [27], we measure the relative performance using Wang *et al.* [101] as baseline. Without using POS, i.e., (w/o)-POS gives a B-4/C score of $31.1/100$ as compared to with POS, i.e., (w)-POS that gives $31.6/104/5$. (ii) For mouse trace [27], we use [73] to get (w/o)-Trace B-4/C score of $8.1/29.3$ as compared to (w)-Trace score of $24.6/106.5$. This leads to a large relative improvement of $16.5/77.2$ to show the significant contribution of using mouse trace as guiding signal. (iii) For sketch, we follow [20] to use [62] as baseline to get (w/o)-Sketch B-4/C score of $31.8/42.7$. We use cross-attention mechanism in [65] to inject sketch as a guiding signal into our baseline [63] to give a (w)-Sketch score of $42.7/121.6$. This gives a relative improvement of $10.9/16.1$, which shows that sketch as a guiding signal is better than POS and competitive as mouse trace. Hence, we advocate for sketch as a guiding signal to depict saliency since unlike POS [27] or mouse trace [65], sketches are more expressive that can capture artistic interpretation like caricature [39].

B. Modelling more than three modalities

Sec. 4.4 optionally models the modality-agnostic components of sketch or text using the function $\mathcal{G}(\cdot)$ that consists of a multihead cross-attention module $\texttt{MH}(\cdot)$ followed by an attention-based pooling $\texttt{PMA}(\cdot)$. For $M = 3$, $\mathcal{L}_{cls}^{tot}$ is defined as,

$$\begin{aligned}\mathcal{L}_{cl}^{tot} =& \mathcal{L}_{cl}(\mathcal{G}(f_{\mathbf{s}}^{ag}, f_{\mathbf{t}}^{ag}), f_{\mathbf{p}}^{ag}) \\ &+ \mathcal{L}_{cl}(\mathcal{G}(f_{\mathbf{s}}^{ag}, f_{\mathbf{p}}^{ag}), f_{\mathbf{t}}^{ag}) + \mathcal{L}_{cl}(\mathcal{G}(f_{\mathbf{p}}^{ag}, f_{\mathbf{t}}^{ag}), f_{\mathbf{s}}^{ag})\end{aligned} \tag{6}$$

In this section, we show how $\mathcal{G}(\cdot)$ can be extended to more than three modalities $M > 3$. Given a set of modality-agnostic components as $\Psi = \{f_{\mathbf{1}}^{ag}, f_{\mathbf{2}}^{ag}, \dots, f_{\mathbf{M}}^{ag}\}$, we can solve for $\mathcal{L}_{cl}^{tot}$ as,

$$\mathcal{L}_{cl}^{tot} = \sum_{j=1}^{\mathbf{M}} \mathcal{L}_{cl}(\mathcal{G}(\Psi - \{f_j^{ag}\}), f_j^{ag}) \tag{7}$$

We further elaborate Eq. (7) using Algorithm 1.

Algorithm 1 Compute generalised $\mathcal{L}_{cl}^{tot}$ for $M > 3$

```
Require: P ∈ R^{1×480}                              ▷ Learned weights.
  Ψ = {f_1^ag, f_2^ag ..., f_M^ag}, ∈ R^{M×480}
  L_cl^tot ← 0
  for j ← 1 to M do
      S_M ← Ψ − {f_i^ag}                            ▷ (M − 1) × 480
      H_M ← MH(S_M)                                 ▷ (M − 1) × 480
      f_M = PMA(H_M) = σ(P H_M^T) H_M               ▷ (1 × 480)
      L_cl^tot ← L_cl^tot + L_cl(f_j^ag, f_M)
  end for
  return L_cl^tot
```

C. Derivation of Disentanglement Loss in Eq. 3

For optionality across tasks, we disentangle the information from sketch, text, and photo, given by $\mathbf{k} \in \{\mathbf{s}, \mathbf{t}, \mathbf{p}\}$ into a discriminative part $f_{\mathbf{k}}^{ag}$ shared across modalities, and a generative part specific to one modality $f_{\mathbf{k}}^{sp}$. This information split of $f_{\mathbf{k}} = [f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp}]$ is achieved in Sec. 4.3 by minimising the mutual information between the modality-agnostic and modality-specific components defined as,

$$\begin{aligned}\mathcal{I}(f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp}) &= \int_{f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp}} \mathbb{p}(f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp}) \log \frac{\mathbb{p}(f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp})}{\mathbb{p}(f_{\mathbf{k}}^{ag})\mathbb{p}(f_{\mathbf{k}}^{sp})} \\ &= \int_{f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp}} \mathbb{p}(f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp}) \log \frac{\mathbb{p}(f_{\mathbf{k}}^{sp} | f_{\mathbf{k}}^{ag})}{\mathbb{p}(f_{\mathbf{k}}^{sp})}\end{aligned} \tag{8}$$

Given a variational distribution $\mathbb{q}(f_{\mathbf{k}}^{sp})$, due to positivity of KL divergence we have,

$$\int \mathbb{p}(f_{\mathbf{k}}^{sp}) \log \mathbb{p}(f_{\mathbf{k}}^{sp}) \geq \int \mathbb{p}(f_{\mathbf{k}}^{sp}) \log \mathbb{q}(f_{\mathbf{k}}^{sp}) \tag{9}$$

Hence, approximating the modality-specific prior $\mathbb{p}(f_{\mathbf{k}}^{sp})$ with variational distribution $\mathbb{q}(f_{\mathbf{k}}^{sp})$ in Eq. (8) we get,

$$\mathcal{I}(f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp}) \leq \int_{f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp}} \mathbb{p}(f_{\mathbf{k}}^{ag}, f_{\mathbf{k}}^{sp}) \log \frac{\mathbb{p}(f_{\mathbf{k}}^{sp} | f_{\mathbf{k}}^{ag})}{\mathbb{q}(f_{\mathbf{k}}^{sp})} \tag{10}$$

Assuming a uniform prior distribution $\mathbb{p}(\eta)$, and its definition in Eq. 2 via conditional invertible neural network $\tau_{\mathbf{k}}$, we have,

$$\begin{aligned}\mathcal{L}_{\tau_{\mathbf{k}}} =& -\mathbb{E}_{f_{\mathbf{k}}^{sp}, f_{\mathbf{k}}^{ag}} \{\log \mathbb{q}(\tau_{\mathbf{k}}^{-1}(f_{\mathbf{k}}^{sp} \mid f_{\mathbf{k}}^{ag})) \\ &+ \log |\det J_{\tau_{\mathbf{k}}^{-1}}(f_{\mathbf{k}}^{sp} \mid f_{\mathbf{k}}^{ag})|\} - H(f_{\mathbf{k}}^{sp} | f_{\mathbf{k}}^{ag})\end{aligned} \tag{11}$$

where, $H(f_{\mathbf{k}}^{sp}|f_{\mathbf{k}}^{ag})$ is the constant data entropy which is ignored in the final optimisation in Eq. 3.

D. Comparison with a parallel work [85]

A parallel work surfaced while writing this paper by Sangkloy *et al*. [85] can optionally perform text-based image retrieval (TBIR), sketch-based image retrieval (SBIR), or both sketch+text based image retrieval (STBIR). However, the motivation of [85] is crucially different from ours. While we focus on improving the latent space via disentanglement into a modality-specific and modality-agnostic component to support optionality across tasks (retrieval and captioning) and modalities (using only sketch, only text, or both as query), Sangkloy *et al*. [85] focused on improving the encoders for sketch, text, and photo by adapting the recently popular pre-trained CLIP [75]. To model only sketch, only text, or both sketch+text for image retrieval, [85] used a rather simple late-fusion technique performing element-wise addition of sketch and text features. While the training code of the proposed model in [85] is not been released yet, our re-implementation of [85] using simple element-wise addition of sketch and text features with CLIP encoders lead to STBIR performance of $23.9/53.5$ in Acc.@1/Acc.@10 which is significantly lower than our proposed method by $15.6/35.2$ on FS-COCO [20]. Although CLIP [75] is highly generalisable to open-set setups, it is difficult to adapt to small downstream datasets like FS-COCO [20] and simultaneously outperform task-specific encoders like VGG-16 [88] used in the proposed method. A similar trend was also observed in Chowdhury *et al*. [20].

E. Clarification on Contributions

Our goal is not to design a model that is state-of-the-art for ALL retrieval (e.g., FG-STBIR, FG-SBIR, FG-TBIR) and generative (e.g., image, sketch, and subjective captioning) tasks. Instead, we (i) design a generalisable model that is competitive with a myriad of baselines (large models like CLIP-LN or small ones like VGG) across multiple tasks; (ii) we show how the benefits of sketch modality (acknowledged by several prior works [20, 96]) can be optionally combined with multiple modalities like text and photo.

F. Comparison with Matrix Factorization

While our baseline MulCap performs feature multiplication similar to matrix factorization [67, 100], we additionally adopt [100] to get subjective captioning (BELU-1, CIDEr) score of (79.2 ± 0.6, 113.5 ± 1.1).

G. Evaluation with different training seeds

Training on 5 different seeds, we report accuracy on FG-STBIR task. For FS-COCO [20] we get Acc.@1 and Acc.@10 of 25.6 ± 0.5 and 55.3 ± 0.3 respectively. Further experimenting on shoe dataset [112], we get FG-STBIR Acc.@1 and Acc.@10 scores of 53.2 ± 0.5 and 88.1 ± 0.2.

H. Additional Details on Pilot Study

Our pilot study aims to: (i) compare sketch vs. text as a query for fine-grained image retrieval. For this, we use standard baselines Triplet-SN (for SBIR) and CLIP-LN (for TBIR) on 3000 sketch/photo, and text/photo pairs in FS-COCO [20]. We observe that for some instances sketch is a better query for image retrieval as it can depict complex shapes, multiple objects, and spatial alignment. However, not all objects are easy to draw (e.g., differentiate a 'donkey' vs. a 'horse') but could be easily described via text. (ii) For subjective captioning, we compare the relative improvements in standard captioning metrics (like M, R, C, S) when using users' sketch (to generate subjective captions) vs. without using sketches (to generate subjective captions).

I. Comparison with Aytar *et al*. [4]

Ayatar *et al*. [4] learns a joint embedding space across image, sound, and text. This is similar to our method, which also aims to learn a joint embedding space across image, sketch, and text. However, there are some key differences: (i) [4] lacks the ability to combine multiple modalities like sound+text for image retrieval. The ability to optionally combine multiple modalities for image retrieval is crucial to our motivation, e.g., fine-grained sketch-based image retrieval (FG-SBIR), fine-grained text-based image retrieval (FG-TBIR), and fine-grained sketch+text based image retrieval (FG-STBIR). (ii) The embedding space of [4] only supports discriminative tasks. This fails to support the generative objectives of our method, like image captioning, sketch captioning, and subjective captioning. Nevertheless, we compare Acc.@1 with [4] on FS-COCO [20] for FG-SBIR and FG-TBIR to get 23.5% and 7.1% respectively.

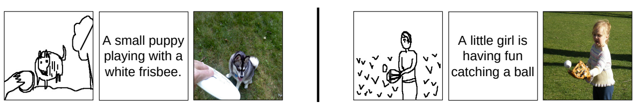
J. Differences from prior works

Prior works like (i) Aytar *et al*. [3] study only cross-modal transfer between a pair of modalities (sketch/photo, or text/photo), not a combination of multiple modalities (sketch+text, or sketch+photo) nor feature disentanglement (modality-agnostic and modality-specific) which is crucial for tasks like FG-STBIR and subjective captioning. (ii) Song *et al*. [89] combines sketch+text for image retrieval via a weighted sum of sketch-photo and text-photo distances computed independently. This simple setup is (a) limited to retrieval (i.e., no captioning), and (b) lacks feature disentanglement to filter our irrelevant modality-specific information (drawing style) when combining multiple modalities (sketch+text). We bring new insights into scene understanding by showing the need for feature disentanglement to

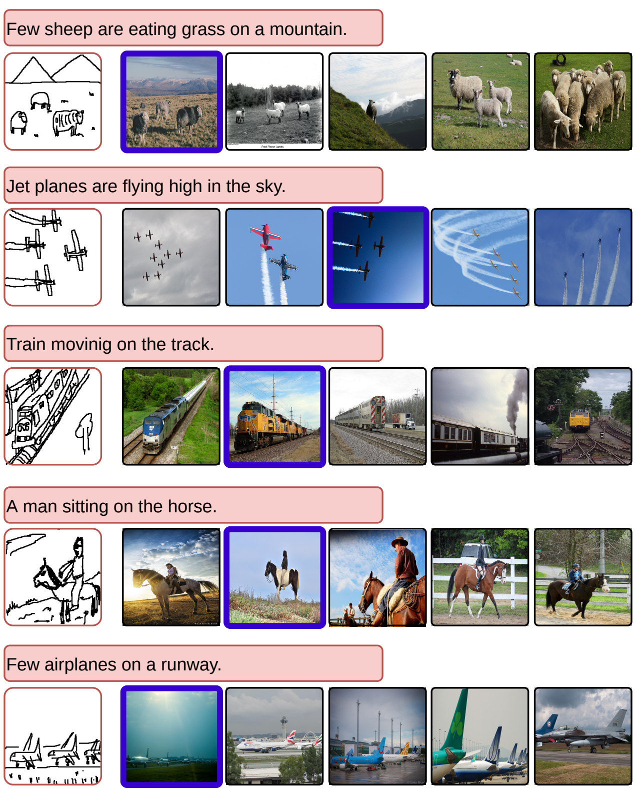
(i) optionally combine multiple modalities, and (ii) support both discriminative and generative tasks.

K. Complex Faliure Cases

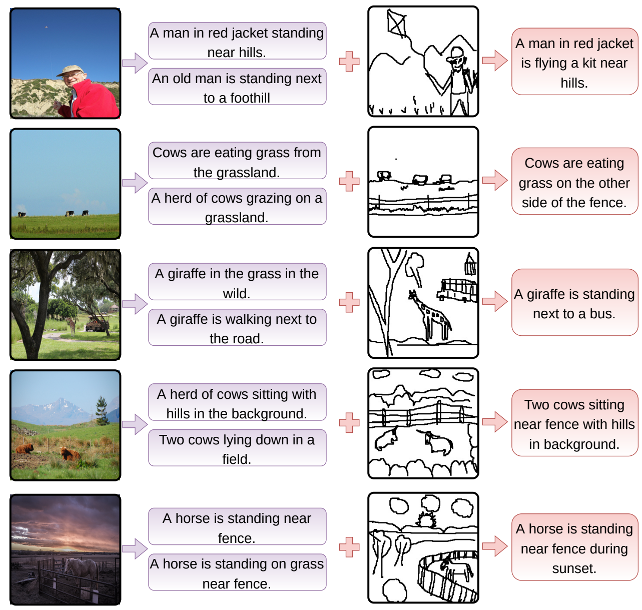
We show qualitative results below where sketch + text performs poorly. We observe this happens when both the input sketch or text is ambigious (i.e., badly drawn sketch or unprecise short textual phrases).



L. Sketch+Text as Query for Image Retrieval



M. Image vs. Subjective Captioning



References

[1] Harsh Agarwal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, 2019. 3

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: semantic propositional image caption evaluation. In *ECCV*, 2016. 6

[3] Yusuf Aytar, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Cross-modal scene networks. *IEEE TPAMI*, 2018. 1, 2, 3, 7, 8, 10

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, Hear, and Read: Deep Aligned Representations. *arXiv preprint arXiv:1706.00932*, 2017. 10

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 6

[6] Zechen Bai, Yuta Nakashima, and Noa Gracia. Explain me the painting: Multi-topic knowledgeable art description generation. In *ICCV*, 2021. 3

[7] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In *CVPR*, 2021. 2

[8] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, and Yi-Zhe Song. Vectorization and rasterization: Self-supervised learning for sketch and handwriting. In *CVPR*, 2021. 2, 3

[9] Ayan Kumar Bhunia, Ayan Das, Umar Riaz Muhammad, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, Yulia Gryaditskaya, and Yi-Zhe Song. Pixelor: A competitive sketching ai agent. so you think you can beat me? *ACM TOG*, 2020. 2

[10] Ayan Kumar Bhunia, Subhadeep Koley, Abdullah Faiz Ur Rahman Khilji, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketching without worrying: Noise-tolerant sketch-based image retrieval. In *CVPR*, 2022. 2

[11] Ayan Kumar Bhunia, Subhadeep Koley, Amandeep Kumar, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketch2Saliency: Learning to Detect Salient Objects from Human Drawings. In *CVPR*, 2023. 2

[12] Ayan Kumar Bhunia, Aneeshan Sain, Parth Shah, Animesh Gupta, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Adaptive fine-grained sketch-based image retrieval. In *ECCV*, 2022. 1, 2

[13] Ayan Kumar Bhunia, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch based image retrieval. In *CVPR*, 2020. 2

[14] Ll. Castrejón, Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba. Learning aligned cross-modal representations from weakly aligned data. In *CVPR*, 2016. 1, 2

[15] Soravit Changpinyo, Bo Pang, Piyush Sharma, and Radu Soricut. Decoupled box proposal and featurization with ultrafine-grained semantic labels improve image captioning and visual question answering. In *EMNLP*, 2019. 7, 8

[16] Long Chen, Zhihong Jiang, Jun Xiao, and Wei Liu. Human-like controllable image captioning with verb-specific semantic roles. In *CVPR*, 2021. 3

[17] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: interpretable representation learning by information maximizing generative adversarial nets. In *NeurIPS*, 2016. 3

[18] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Viswanatha Reddy Gajjala, Aneeshan Sain, Tao Xiang, and Yi-Zhe Song. Partially Does It: towards scene-level FG-SBIR with partial input. In *CVPR*, 2022. 2

[19] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. What Can Human Sketches Do for Object Detection? In *CVPR*, 2023. 2

[20] Pinaki Nath Chowdhury, Aneeshan Sain, Yulia Gryaditskaya, Ayan Kumar Bhunia, Tao Xiang, and Yi-Zhe Song. Fs-coco: Towards understanding of freehand sketches of common objects in context. In *ECCV*, 2022. 1, 2, 3, 6, 7, 8, 9, 10

[21] Pinaki Nath Chowdhury, Tuanfeng Wang, Duygu Ceylan, Yi-Zhe Song, and Yulia Gryaditskaya. Garment ideation: Iterative view-aware sketch-based garment modeling. In *3DV*, 2022. 2

[22] Sanghyuk Chun, Joon Oh, Seong, Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *CVPR*, 2021. 2

[23] John Collomosse, Tu Bui, and Jin Hailin. Livesketch: Query perturbations for guided sketch-based visual search. In *CVPR*, 2019. 2

[24] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, 2020. 1

[25] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[26] Michael J. Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *WMT@ACL*, 2014. 6

[27] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander Schwing, and D. A. Forsyth. Fast, diverse and accurate image captioning guided by part-of-speech. In *CVPR*, 2019. 1, 3, 9

[28] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *CVPR*, 2019. 2

[29] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. In *ICLR*, 2015. 4

[30] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estiamtion using real nvp. In *ICLR*, 2017. 6

[31] Alexey Dosovitskiy, Lucas. Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold,

Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is wort 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 6, 8

[32] Aviv Eisenschtat and Loir Wolf. Linking image and text with 2-way nets. In *CVPR*, 2017. 2

[33] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM TOG*, 2012. 1

[34] Fartash Faghri, David J. Fleet, Jaime Ryan Kiros, and Sanja Fidler. Vse++: Improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612*, 2017. 2

[35] Chengying Gao, Qi Liu, Limin Wang, Jianzhuang Liu, and Changqing Zou. Sketchycoco: Image generation from freehand scene sketches. In *CVPR*, 2020. 1, 2, 6, 7, 8

[36] Lluís Gómez, Andrés Mafla, Marçal Rusiñol, and Dimosthenis Karatzas. Single shot scene text retrieval. In *ECCV*, 2018. 1

[37] Yuyu Guo, Xuanhan Gao, Liali ad Wang, Yuxuan Hu, Xing Xu, Xu Lu, Heng Tao Shen, and Jingkuan Song. From general to specific: Informative scene graph generation via balance adjustment. In *ICCV*, 2021. 3

[38] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *ECCV*, 2020. 3

[39] Xiaoguang Han, Kangcheng Hou, Dong Du, Yuda Qiu, Yizhou Yu, Kun Zhou, and Shugang Cui. Caricatureshop: Personalized and photorealistic caricature sketching. *IEEE TVCG*, 2018. 3, 8, 9

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[41] Weixiang Hong, Kaixiang Ji, Jiajia Liu, Jian Wang, Jingdong Chen, and Wei Chu. Gilbert: Generative vision-language pre-training for image-text retrieval. In *SIGIR*, 2021. 3

[42] Wei-Ning Hsu and James Glass. Disentangling by partitioning: A representation learning framework for multimodal sensory data. *arXiv preprint arXiv:1805.11264*, 2018. 3

[43] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 3

[44] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2, 6

[45] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 3

[46] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018. 4, 6

[47] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 6

[48] Kazuma Kobayashi, Lin Gu, Ryuichiro Hataya, Takaaki Mizuno, Mototaka Miyake, Hirokazu Watanabe, Masamichi Takahashi, Yasuyuki Takamizawa, Yukihiro Yoshida, Satoshi Nakamura, Nobuji Kouno, Amina Bolatkan, Yusuke Kurose, Tatsuya Harada, and Ryuji Hamamoto. Sketch-based Medical Image Retrieval. *arXiv preprint arXiv:2303.03633*, 2023. 2

[49] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Picture that Sketch: Photorealistic Image Generation from Abstract Sketches. In *CVPR*, 2023. 2

[50] Iro Laina, Christian Rupprecht, and Nassir Navab. Towards unsupervised image captioning with shared multimodal embeddings. In *ICCV*, 2019. 3

[51] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *ICML*, 2019. 2

[52] Lei Li, Changqing Zou, Youyi Zheng, Qingkun Su, Honbo Fu, and Chiw-Lan Tai. Sketch-r2cnn: An attentive network for vector sketch recognition. *arXiv preprint arXiv:1811.08170*, 2018. 3

[53] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 2

[54] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004. 6

[55] Hangyu Lin, Yanwei Fu, Yu-Gang Jiang, and Xiangyang Xue. Sketch-bert: Learning sketch bidirectional encoder representation from transformers by self-supervised learning of sketch gestalt. In *CVPR*, 2020. 3

[56] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: common objects in context. In *ECCV*, 2014. 1, 3, 6, 8

[57] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *CVPR*, 2021. 3

[58] Fang Liu, Changqing Zhou, Xiaoming Deng, Ran Zuo, Yu-Kun Lai, Cuixia Ma, Yong-Jin Liu, and Hongan Wang. Scenesketcher: Fine-grained image retrieval with scene sketches. In *ECCV*, 2020. 1, 2, 3, 6, 7

[59] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to combine modalities in multimodal deep learning. *arXiv preprint arXiv:1805.11730*, 2018. 5, 7

[60] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *ECCV*, 2018. 1, 4

[61] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 3

[62] Shweta Mahajan, Iryna Gurevych, and Stefan Roth. Latent normalizing flows for many-to-many cross-domain mappings. In *ICLR*, 2020. 1, 7, 8, 9

[63] Shweta Mahajan and Stefan Roth. Diverse image captioning with context-object split latent spaces. In *NeurIPS*, 2020. 1, 3, 9

[64] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In *CoLT*, 2021. 2

[65] Zihang Meng, Licheng Yu, Ning Zhang, Tamara Berg, Babak Damavandi, Vikas Singh, and Amy Bearman. Connecting what to say with where to look by modeling human attention traces. In *CVPR*, 2021. 3, 9

[66] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 7

[67] Nils Murrugarra-Llerena and Adriana Kovashka. Cross-Modality Personalization for Retrieval. In *CVPR*, 2019. 10

[68] Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Cross-domain generative learning for fine-grained sketch-based image retrieval. In *BMVC*, 2017. 2

[69] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 2, 6

[70] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *CVPR*, 2017. 3

[71] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Towards personalized image captioning via multimodal memory networks. *IEEE TPAMI*, 2018. 3

[72] Bryan A. Plummer, Paige Kordas, M. Hadi Kaipour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. Conditional image-text embedding networks. *arXiv preprint arXiv:1711.08389*, 2017. 2

[73] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 3, 7, 9

[74] Anran Qi, Yulia Gryaditskaya, Tao Xiang, and Yi-Zhe Song. One sketch for all: One-shot personalized sketch segmentation. *TIP*, 2022. 2

[75] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2, 3, 6, 7, 8, 10

[76] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019. 7

[77] Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *ICML*, 2014. 3

[78] Leo Sampaio Ferraz Riberio, Tui Bui, John Collomosse, and Moacir Ponti. Scene designer: a unified model for scene search and synthesis from sketch. In *ICCV Workshop*, 2021. 2, 3

[79] Paul K. Rubenstein, Bernhard Schoelkopf, and Ilya Tolstikhin. On the latent space of wasserstein auto-encoders. *arXiv preprint arXiv:1802.03761*, 2018. 3

[80] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. CLIP for All Things Zero-Shot Sketch-Based Image Retrieval, Fine-Grained or Not. In *CVPR*, 2023. 2

[81] Aneeshan Sain, Ayan Kumar Bhunia, Subhadeep Koley, Pinaki Nath Chowdhury, Soumitri Chattopadhyay, Tao Xiang, and Yi-Zhe Song. Exploiting Unlabelled Photos for Stronger Fine-Grained SBIR. In *CVPR*, 2023. 2

[82] Aneeshan Sain, Ayan Kumar Bhunia, Vaishnav Potlapalli, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketch3t: Test-time training for zero-shot sbir. In *CVPR*, 2022. 2

[83] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Cross-modal hierarchical modelling for fine-grained sketch based image retrieval. In *BMVC*, 2020. 6

[84] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *CVPR*, 2021. 3, 4

[85] Patsorn Sangkloy, Wittawat Jitkrittum, Diyi Yang, and James Hays. A sketch is worth a thousand words: Image retrieval with text and sketch. In *ECCV*, 2022. 10

[86] Vaswani A. Shazeer, N. Parmar, N. Uszkoreit, J. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 6, 7

[87] Kurt Shuster, Samuel Humeau, Hexiang Hu, Antonie Bordes, and Jason Weston. Engaging image captioning via personality. In *CVPR*, 2019. 3

[88] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 6, 10

[89] Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Fine-grained image retrieval: the text/sketch input dilemma. In *BMVC*, 2017. 3, 7, 8, 10

[90] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*, 2017. 6, 7

[91] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, 2018. 4

[92] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. *arXiv preprint arXiv:2208.04254*, 2021. 3, 8

[93] Joshua Susskind, Adam Anderson, and Geoffrey Hinton. The toronto face dataset. Technical report, Toronto University, 2010. 3

[94] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. In *ICLR Workshop*, 2017. 3

[95] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 3

[96] Aditay Tripathi, Rajath R. Dani, Anand Mishra, and Anirban Chakraborty. Sketch-guided object localization in natural images. In *ECCV*, 2020. 10

[97] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *ICLR*, 2019. 3

[98] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5, 8

[99] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 2, 6

[100] Andreas Veit, Maximilian Nickel, Serge Belongie, and Laurens van der Maaten. Separating Self-Expression and Visual Content in Hashtag Supervision. In *CVPR*, 2018. 10

[101] Liwei Wang, Alexander G. Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *NeurIPS*, 2017. 3, 6, 7, 8, 9

[102] Xi Wang, Kathleen Ang, and Faramarz Samavati. Sketch-based editing and deformation of cardiac image segmentation, 2022. 2

[103] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *ICCV*, 2019. 2

[104] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *CSCW*, 2017. 3

[105] Jun Xing, Li-Yi Wei, Takaaki Shiratori, and Koji Yatani. Autocomplete hand-drawn animations. *ACM TOG*, 2015. 2

[106] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 1, 4, 6, 8

[107] Peng Xu, Timothy M. Hospedales, Qiyue Yin, Yi-Zhe Song, Tao Xiang, and Liang Wang. Deep learning for free-hand sketch: A survey. *IEEE TPAMI*, 2022. 2

[108] Peng Xu, Xiatian Zhu, and David A. Clifton. Multi-modal learning with transformers: A survey. *arXiv preprint arXiv:2206.06488*, 2022. 3

[109] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing inter-modality: Visual parsing with self-attention for vision-language pre-training. In *NeurIPS*, 2021. 3

[110] Shuai Yang, Zhangyang Wang, Jiaying Liu, and Zongming Guo. Deep plastic surgery: Robust and controllable image editing with human-drawn sketches. In *ECCV*, 2020. 2

[111] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, 2018. 2

[112] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen Change Loy. Sketch me that shoe. In *CVPR*, 2016. 2, 3, 6, 7, 10

[113] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-net: A deep neural network that beats humans. *IJCV*, 2017. 6

[114] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 6

[115] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xialei Huang, and Dimitris N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. In *IEEE TPAMI*, 2019. 6

[116] Wei Zhang, Yue Ying, Pang Lu, and Hongyuan Zha. Learning long- and short-term user literal-preference with multimodal hierarchical transformer network for personalized image caption. In *AAAI*, 2020. 3

[117] Youyan Zhang, Jiuniu Wang, Hao Wu, and Wenjia Xu. Distinctive image captioning via clip guided group optimization. *arXiv preprint arXiv:2208.04254*, 2022. 3

[118] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Olivia, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 2017. 1

[119] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019. 1

[120] Changqing Zou, Qian Yu, Ruofei Du, Haoran Mo, Yi-Zhe Song, Tao Xiang, Chengying Gao, Baoquan Chen, and Hao Zhang. Sketchyscene: Richly-annotated scene sketches. In *ECCV*, 2018. 1, 2, 6, 7