

# SkySearch: Satellite Video Search at Scale (Online Appendix)

Minyoung Choe<sup>\*1</sup>, Geon Lee<sup>\*1</sup>, Changhun Han<sup>\*2</sup>, Suji Kim<sup>2</sup>, Woong Hu<sup>3</sup>, Hyebeen Hwang<sup>3</sup>,  
Geunseok Park<sup>3</sup>, Byeongyeon Kim<sup>4</sup>, Hyesook Lee<sup>4</sup>, Ha-Myung Park<sup>†2</sup>, and Kijung Shin<sup>†1</sup>

<sup>1</sup> Kim Jaechul Graduate School of AI, KAIST, Seoul, South Korea, {minyoung.choe, geonlee0325, kijungs}@kaist.ac.kr

<sup>2</sup> College of Computer Science, Kookmin University, Seoul, South Korea, {codingnoye, suji2924, hmpark}@kookmin.ac.kr

<sup>3</sup> Satrec Initiative Co., Ltd., Daejeon, South Korea, {woonghu, hbh, gspark}@satreci.com

<sup>4</sup> National Institute of Meteorological Sciences, Jeju, South Korea {bykim1011, hslee05}@korea.kr

## I. METHOD DETAILS

We present additional details of SKYSEARCH that are not included in the main paper due to space constraints.

**Training details on data compression.** We detail the training process for the data compression module in SKYSEARCH, which consists of an image encoder and a video encoder. The image encoder captures spatial information using convolutional layers, while the video encoder captures temporal information by processing sequences of image embeddings through a Convolutional RNN. Training is conducted in two stages: first, the image encoder is trained independently, and then the video encoder is trained while keeping the image encoder frozen. Both encoders are optimized using a self-supervised loss function. For a given image (or video)  $v$ , we define  $P_v$  as the set of positive images (or videos) that are temporally close to  $v$ , and  $N_v$  as the set of negative images (or videos) that are temporally distant:

$$P_v = \{v' : |t_v - t_{v'}| \leq \Delta\}, \quad N_v = \{v' : |t_v - t_{v'}| > \Delta\}$$

where  $\Delta$  is a predefined temporal threshold, set to  $\Delta = 8$  hours by default. And  $t_v$  denotes the timestamp of the image or the initial timestamp of the video. The self-supervised loss is defined as:

$$\mathcal{L}_v = \mathbb{E}_{\substack{p \sim P_v \\ n \sim N_v}} \max(\|f(v) - f(p)\|_2^2 - \|f(v) - f(n)\|_2^2 + \gamma, 0)$$

Here  $f(\cdot)$  represents the image or video encoder, and we use  $\gamma$  as 0.5. Equal numbers of positive and negative samples are used during training. The final loss aggregates all images or videos in the database, i.e.,  $\mathcal{L} = \sum_v \mathcal{L}_v$ .

**Training details on video prediction.** We provide further details about the video prediction module of SKYSEARCH. While SimVP [1] is trained with a pixel-wise MSE loss to predict future video frames, it fails to generate high-quality outputs when applied to high-dimensional satellite videos, as shown in Figures 5 and 6. To address this limitation, SKYSEARCH introduces an adversarial loss by incorporating a discriminator  $D$  in addition to the generator (i.e., video predictor)  $G$ . Specifically, given an input video  $v$  and its

ground-truth future video  $v'$ , a discriminator  $D$  is trained to distinguish between the real future video  $v'$  and the generated estimated future video  $G(v)$ . The adversarial loss for the discriminator is defined as:

$$\mathcal{L}_{\text{adv}}^{(D)} = -\mathbb{E}[\log D(v')] - \mathbb{E}[\log(1 - D(G(v)))].$$

The generator  $G$  is trained to minimize two losses. The first loss is the MSE loss, defined as:

$$\mathcal{L}_{\text{MSE}} = \|v' - G(v)\|_2^2,$$

which aims to make the generated video  $G(v)$  closely match the ground-truth future video  $v'$  at the pixel level. The second is the adversarial loss, defined as:

$$\mathcal{L}_{\text{adv}}^{(G)} = -\mathbb{E}[\log(D(G(v)))],$$

which encourages the generator to generate video frames that are indistinguishable from real frames by the discriminator.

We further introduce two auxiliary losses to enhance the prediction quality. For the generator, we introduce an orthogonal regularization loss that aims to promote the orthogonality of its parameters, helping to improve generalization and prevent the mode collapse issue. This loss is defined as:

$$\mathcal{L}_{\text{orth}} = \sum_i \|W_i W_i^T - I\|_F^2,$$

where  $W_i$  is the weight matrix of the  $i$ -th layer in the generator, and  $I$  is the identity matrix. For the discriminator, we introduce a feature reconstruction loss to align the latent features of the ground-truth future video  $v'$  and the generated video  $G(v)$ . This loss encourages the discriminator to better align the generated video with the actual video. It is defined as:

$$\mathcal{L}_{\text{feat}} = \|\phi(v') - \phi(G(v))\|_2^2,$$

where  $\phi(\cdot)$  is the feature extracted from the discriminator.

To effectively train SKYSEARCH, it is important to balance the contribution of the three main losses. We introduce three hyperparameters,  $\lambda_{\text{MSE}}$ ,  $\lambda_{\text{adv}}^{(G)}$  and  $\lambda_{\text{adv}}^{(D)}$ , which control the weighting of the MSE loss, the adversarial loss for the generator, and the adversarial loss for the discriminator, respectively. These weights are normalized such that  $\lambda_{\text{MSE}} + \lambda_{\text{adv}}^{(G)} + \lambda_{\text{adv}}^{(D)} = 1$ . At each training step, we perform biased sampling to select one of the three losses,  $\mathcal{L}_{\text{MSE}}$ ,  $\mathcal{L}_{\text{adv}}^{(G)}$ , or  $\mathcal{L}_{\text{adv}}^{(D)}$  with probabilities

<sup>\*</sup> Co-first authors.

<sup>†</sup> Co-corresponding authors.

proportional to  $\lambda_{\text{MSE}}$ ,  $\lambda_{\text{adv}}^{(G)}$  and  $\lambda_{\text{adv}}^{(D)}$ , respectively, to minimize. The two auxiliary loss terms,  $\mathcal{L}_{\text{orth}}$  and  $\mathcal{L}_{\text{feat}}$ , are optimized when the generator or the discriminator are chosen to be trained in the respective training steps. Empirically, we find that  $\lambda_{\text{MSE}} = 0.3$ ,  $\lambda_{\text{adv}}^{(G)} = 0.6$  and  $\lambda_{\text{adv}}^{(D)} = 0.1$  provide a good balance for effective training.

## II. ADDITIONAL EMPIRICAL RESULTS

We present additional experimental results of SKYSEARCH.

**Extra similar video search results.** As shown in the numerical results in the main paper, SKYSEARCH consistently outperforms ResNet and EfficientNet (both pre-trained and fine-tuned) in both short-term and long-term accuracy. Notably, SKYSEARCH with video prediction achieves the highest accuracy for long-term retrieval, while SKYSEARCH without video prediction excels in short-term accuracy. To complement these numerical results with qualitative evidence, we present additional long-term search results comparing SKYSEARCH to the fine-tuned EfficientNet, the best-performing baseline. To provide a more detailed view of the long-term (24-hour) retrieval, four representative frames from each retrieved 24-hour video are included. As illustrated in Figures 1 and 2, SKYSEARCH with video prediction retrieves videos that are more similar to the query and its future developments compared to both the baseline and SKYSEARCH without video prediction. These results demonstrate the unique ability of SKYSEARCH with video prediction to retrieve videos that not only align with the query but also capture its anticipated future evolution.

**Extra results with alternative ranking methods.** In SKYSEARCH, once the set  $C$  of similar video candidates is retrieved, they are ranked based on their embedding distances to the query embedding in the latent space by default. For enhanced precision, the ranking can be refined using more computationally intensive similarity or distance measures, such as LPIPS [2], FSIM [3], and SSIM [2]. To illustrate this refinement, we provide additional search results using LPIPS, FSIM, and SSIM to select the best match among 50 similar video candidates retrieved by SKYSEARCH with the video prediction module. As shown in Figures 3 and 4, all the retrieved videos are visually similar to the query video. However, results can be further refined using LPIPS, FSIM, and SSIM for improved precision. It is important to note that while embedding distances are computationally efficient, they may result in less precise retrievals compared to these more advanced measures. On the other hand, refining results with LPIPS, FSIM, and SSIM requires significant computational resources, including loading high-resolution images and performing detailed similarity calculations.

**Extra video prediction results.** In SKYSEARCH, video prediction is used for query video augmentation and visualization. In Figures 5 and 6, we present additional prediction results comparing the video prediction module in SKYSEARCH with the baseline SimVP [1]. SKYSEARCH’s video prediction mod-

ule, enhanced with adversarial learning, yields more accurate and sharper video predictions.

**Extra candidate search performance results.** Figure 7 shows the queries per second (QPS) under varying query time interval lengths (QILs) for all channels (IR, SWIR, and WV). The number  $|C|$  of candidate is set to 50, 100, and 500. The results demonstrate that SKYSEARCH consistently outperforms the baseline methods PyNNDescent and BSBF regardless of the channels.

## REFERENCES

- [1] Z. Gao, C. Tan, L. Wu, and S. Z. Li, “Simvp: Simpler yet better video prediction,” in *CVPR*, 2022.
- [2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *TIP*, vol. 13, no. 4, pp. 600–612, 2004.
- [3] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “Fsim: A feature similarity index for image quality assessment,” *TIP*, vol. 20, no. 8, pp. 2378–2386, 2011.



Fig. 1: Qualitative comparison of search results (1 / 2). SKYSEARCH retrieves satellite videos that are more similar to the query videos, compared to those retrieved by the best-performing baseline (i.e., fine-tuned EfficientNet). Moreover, augmenting queries with video prediction enhances 24-hour (long-term) search accuracy.



Fig. 2: Qualitative comparison of search results (2 / 2). SKYSEARCH retrieves satellite videos that are more similar to the query videos, compared to those retrieved by the best-performing baseline (i.e., fine-tuned EfficientNet). Moreover, augmenting queries with video prediction enhances 24-hour (long-term) search accuracy.

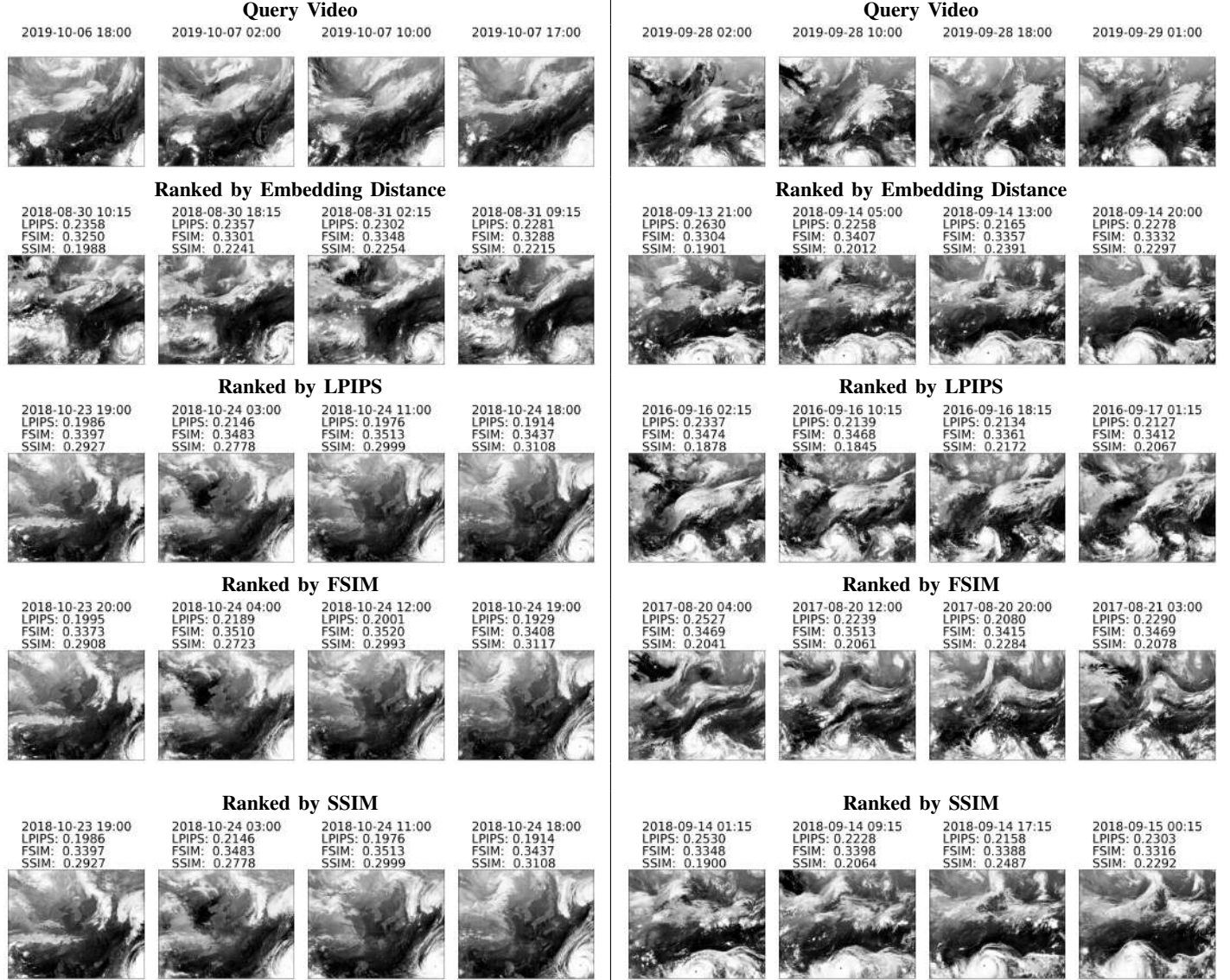


Fig. 3: Results of Alternative Ranking Methods (1 / 2) SKYSEARCH ranks the most similar videos from the candidates retrieved by the candidate search module using embedding distances, LPIPS, SSIM, or FSIM. While all retrieved videos are visually similar to the query, image-based measures such as LPIPS, SSIM, and FSIM can provide further refinement at the cost of higher computational complexity.





Fig. 4: Results of Alternative Ranking Methods (2 / 2) SKYSEARCH ranks the most similar videos from the candidates retrieved by the candidate search module using embedding distances, LPIPS, SSIM, or FSIM. While all retrieved videos are visually similar to the query, image-based measures such as LPIPS, SSIM, and FSIM can provide further refinement at the cost of higher computational complexity.



Fig. 5: Video prediction results (1 / 2). SKYSEARCH employs adversarial learning for video prediction, leading to more accurate and sharper future predictions of the query video compared to SimVP [1].



Fig. 6: Video prediction results (2 / 2). SKYSEARCH employs adversarial learning for video prediction, leading to more accurate and sharper future predictions of the query video compared to SimVP [1].



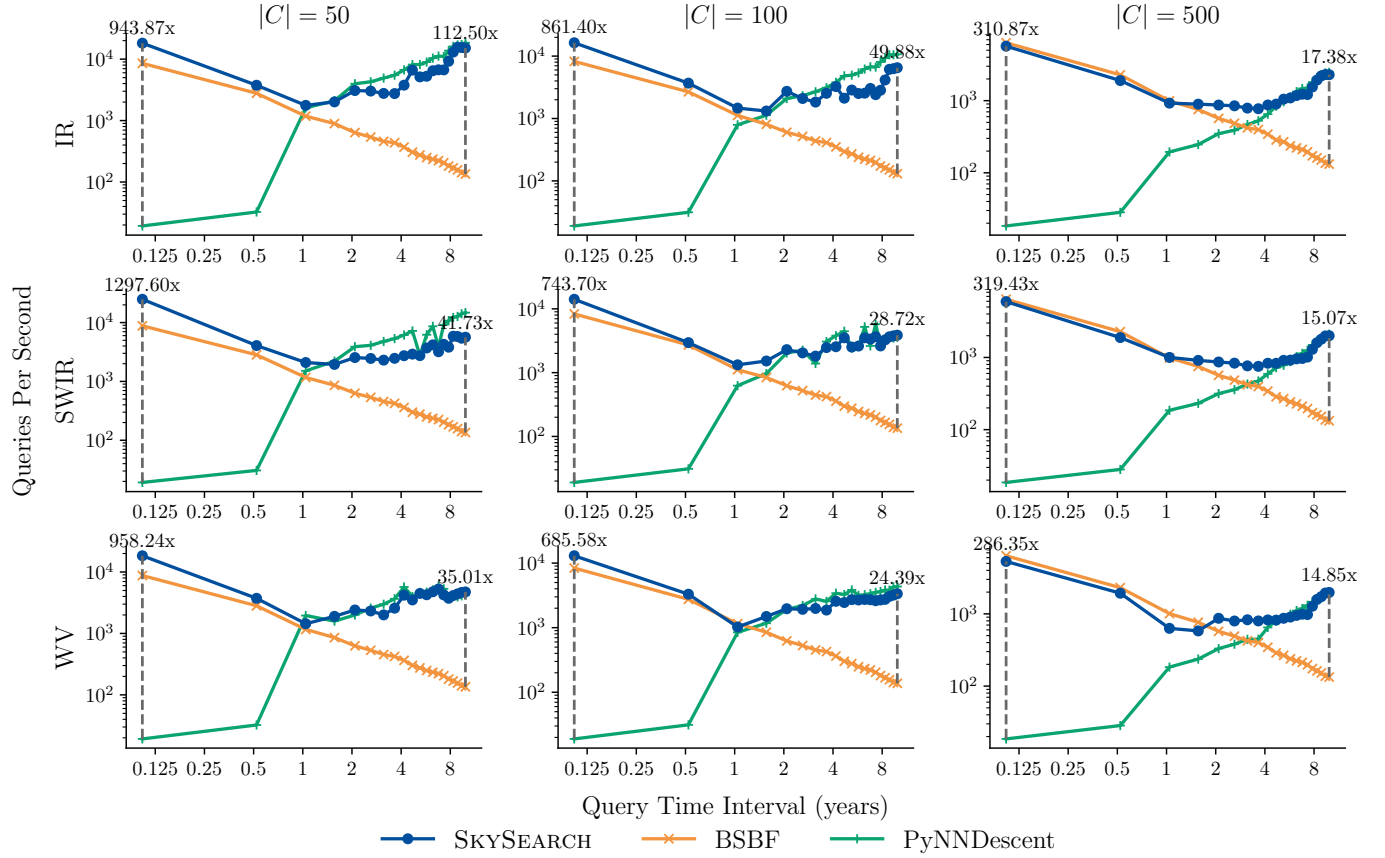


Fig. 7: Candidate search performance for all channels (IR, SWIR, and WV). Regardless of the channels, SKYSEARCH consistently outperforms the baseline methods PyNNDescent and BSBF by up to 1297.60x and 112.50x, respectively.