Can Image Splicing and Copy-Move Forgery Be Detected by the Same Model? Forensim: An Attention-Based State-Space Approach

Soumyaroop Nandi[1,2] Prem Natarajan[1,2]
[1]USC Information Sciences Institute, Marina del Rey, CA, USA
[2]USC Thomas Lord Department of Computer Science, Los Angeles, CA, USA
{soumyarn, premkumn}@usc.edu

Supplementary Material

In this supplementary document, we provide extended technical and experimental details supporting our main paper. Section 1 outlines the Forensim model architecture and implementation details, including the backbone configuration, attention modules, fusion strategy, and training setup. Section 2 presents a comprehensive overview of the datasets used for training and evaluation, including both synthetic and real-world forgery benchmarks. Section 3 provides ablation studies on the loss functions, examining the individual and combined contributions of Cross-Entropy, Dice, Focal, and InfoNCE losses. Section 4 explores the impact of dataset composition on generalization via ablation studies on the proposed `CMFD_Anything` dataset. Finally, Section 5 presents additional robustness analysis on CoMoFoD attacks, highlighting Forensim's performance across diverse manipulation categories.

Index Terms— Copy-Move Forgery Detection (CMFD), Image Manipulation Detection and Localization (IMDL), State Space Models, Attention Mechanisms, Synthetic Datasets.

1. Additional Implementation Details

Backbone and Feature Extraction. Forensim uses the first four layers of a Vision Transformer (ViT) backbone pretrained with DINO on ImageNet. Input images are resized to 224×224 and passed through the backbone to extract hierarchical features $V \in \mathbb{R}^{B\times N\times C}$, where B is batch size, $N = H \times W$, and $C = 384$ is the embedding dimension.

Similarity and Manipulation Attention. The extracted features are processed by two specialized attention modules: the Similarity State Space Attention (Sim_Attn) and Multi-Level Manipulation State Space Attention (MSSA) modules, described in Sections 3.3 and 3.4 of the main text, respectively. The SSA module computes an affinity matrix using state-space recurrence with Rotary Positional Embeddings (RoPE), while the MSSA module integrates manipulation-aware information using multi-head self-attention with Locally Enhanced Positional Encoding (LePE) [5].

Fusion and Prediction. Features from SSA and MSSA are fused via a Non-Local Refinement (NLR) module, which performs global context aggregation based on similarity and manipulation maps. The fused representation is decoded using a lightweight convolutional head with SiLU activation to generate a pixel-level three-class segmentation mask (pristine, source, target) and an image-level detection score.

Loss Functions. We optimize Forensim using a weighted combination of pixel-wise Cross-Entropy Loss (CE), InfoNCE contrastive loss, Dice Loss (DL), and Focal Loss (FL). The overall loss is:

$$\mathcal{L} = \lambda_{\text{CE}} \cdot \mathcal{L}_{\text{CE}} + \lambda_{\text{InfoNCE}} \cdot \mathcal{L}_{\text{InfoNCE}} + \lambda_{\text{DL}} \cdot \mathcal{L}_{\text{Dice}} + \lambda_{\text{FL}} \cdot \mathcal{L}_{\text{Focal}},$$

with default weights: $\lambda_{\text{CE}} = 1.0$, $\lambda_{\text{InfoNCE}} = 0.1$, $\lambda_{\text{DL}} = 1.0$, and $\lambda_{\text{FL}} = 0.5$. Ablations for loss terms are discussed in Section 2.

Contrastive Supervision. Positive pairs are sampled from source-target regions, and negatives from pristine patches. In each batch, we randomly sample 64 positive and 256 negative pairs. Contrastive training encourages discriminative feature learning for manipulated vs. pristine regions.

Training Configuration. Training is performed in PyTorch on an NVIDIA RTX A5000 GPU with 24GB memory. We use the AdamW optimizer with a cyclic learning rate scheduler ranging from 10^{-3} to 10^{-5} and StepLR decay by 0.5 every 10 epochs. Models are trained for 100 epochs with a batch size of 64, sampling 100K images per epoch. Early stopping is applied based on validation loss.

Evaluation Protocol. Models are evaluated using both pixel-level and image-level metrics. Pixel-level performance includes precision, recall, F1, MCC, AUC, and Balanced Accuracy (BAcc), computed using three-class RGB masks. Following prior works [8, 21], we apply a 200-pixel

threshold to suppress small false positives. Image-level scores are derived by averaging segmentation confidence. Additionally, we reported the metric performance for Pristine (P), Source (S) and Target (T) regions for CMFD pixel-level evaluations in Table 2 of the main text.

2. Details on Datasets

Table 1 summarizes the datasets used for training and evaluation. For synthetic CMFD training, we include datasets with copy-move manipulations such as CASIA CMFD, CoMoFoD CMFD, USC-ISI CMFD, and our `CMFD_Anything`, which uniquely includes both copy-move and splicing manipulations synthesized using masks from the Segment Anything model [9]. It is to be noted that our `CMFD_Anything` is the only dataset including pristine images for CMFD training.

We evaluated the Forensim model on three benchmark CMFD datasets—USC-ISI CMFD [20], CoMoFoD [17], and CASIA CMFD [4]—as well as the test split of our proposed `CMFD_Anything` dataset and four baseline IMDL evaluation datasets: NIST16 [1], Columbia [14], Coverage [18], and CASIA [4]. The USC-ISI CMFD dataset consists of 80K training images and 10K each for validation and testing. CoMoFoD contains 5,000 forged images generated from 200 base images across 25 manipulation categories, combining five manipulation types and five post-processing operations. CASIA CMFD includes 1,313 forged and 1,313 authentic images, totaling 2,626 samples.

For IMDL evaluation, we focus on datasets that contain real-world manipulations: CASIA, Columbia, Coverage, and NIST16. These datasets span both splicing and copy-move tasks, offering a robust benchmark for assessing generalization to natural image forensics. Notably, while some datasets like Columbia and Coverage contain only splicing, others such as NIST16 and CASIA include both manipulation types, supporting a comprehensive evaluation across forgery categories.

3. Additional Ablation Study on Forensim Loss Functions

Cross-Entropy Loss for Three Classes

The Cross-Entropy Loss [6] for a multi-class classification problem with three classes (pristine, source, target) is given by:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}) \tag{1}$$

where:

- N is the number of samples (images).
- C is the number of classes (in this case, 3 classes: pristine, source, and target).
- $y_{i,c}$ is the ground-truth label for sample i and class c (binary: 1 if the sample belongs to the class, otherwise 0).
- $p_{i,c}$ is the predicted probability for sample i and class c (obtained from the model's output).

This formula computes the loss by summing over all samples and classes, applying the logarithm to the predicted probabilities, and averaging over the batch size N.

InfoNCE Loss for Three Classes

The InfoNCE loss [15] for a contrastive learning setting with three classes (pristine, source, target) is given by:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathbf{z}_i\cdot\mathbf{z}_i^{+}/\tau)}{\exp(\mathbf{z}_i\cdot\mathbf{z}_i^{+}/\tau)+\sum_{j=1}^{M}\exp(\mathbf{z}_i\cdot\mathbf{z}_j^{-}/\tau)} \tag{2}$$

where:

- N is the number of samples (images).
- $\mathbf{z}_i$ is the feature vector for sample i.
- $\mathbf{z}_i^{+}$ is the feature vector of the positive sample (e.g., source-target pair for i).
- $\mathbf{z}_j^{-}$ is the feature vector of the negative samples (e.g., other regions or non-related images).
- τ is the temperature parameter that controls the smoothness of the softmax.
- M is the number of negative samples (2 in this case, for each positive pair).

This loss function minimizes the distance between the positive pairs while maximizing the distance from the negative pairs, thereby improving the model's ability to distinguish between the different image regions.

Dice Loss

The Dice Loss [13], often used for evaluating segmentation tasks, is given by:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i} \tag{3}$$

where:

- p_i is the predicted probability (or binary prediction) for pixel i.
- g_i is the ground truth (binary) value for pixel i.
- N is the total number of pixels in the image.

The Dice Loss is based on the Dice coefficient, a measure of overlap between two binary sets, and is especially useful when dealing with imbalanced classes. It penalizes differences between the predicted and ground truth segmentation masks by maximizing the overlap between them.

Focal Loss

Focal Loss [12], a modification of cross-entropy loss, is defined as:

Table 1. Summary of image manipulation datasets used for training and evaluation. Datasets are grouped by use case: training on synthetic manipulations, and testing on natural image forensics (IMDL). Manipulation types: splicing and copy-move for natural image forensics.

Dataset Name	Real	Fake	Splicing	Copy-move
Training, Validation, Test (8:1:1) : Synthetic Manipulations (CMFD on Natural Images)				
Casia CMFD [4]	1313	1313	✗	✓
CoMoFoD CMFD [17]	0	5000	✗	✓
Synthetic Images from MSCOCO [11] and SUN2012 [22] for USC-ISI CMFD [20]	0	100K	✗	✓
Synthetic Images from Segment Anything [9] for CMFD_Anything (Ours)	100K	200K	✓	✓
Test: Natural Image Forensics (IMDL)				
CASIA [4]	7,491	5,105	✓	✓
Coverage [18]	100	100	✓	✗
Columbia [14]	183	180	✓	✗
NIST16 [1]	160	160	✓	✓

Table 2. Ablation study on loss functions. Pixel-level metrics on the CMFD_Anything test set that include MCC (Matthews Correlation Coefficient), F1 score (Target class), AUC (Area Under the Curve), and BAcc (Balanced Accuracy). CE = Cross-Entropy, INCE = InfoNCE, DL = Dice Loss, FL = Focal Loss.

Loss Function	MCC [3]	F1 (Target)	AUC	BAcc
CE + DL	0.652	0.600	0.689	0.796
CE + FL	0.638	0.590	0.684	0.782
CE only	0.611	0.572	0.678	0.776
INCE only	0.589	0.548	0.665	0.761
CE + INCE	**0.681**	**0.624**	**0.700**	**0.812**

$$\mathcal{L}_{\text{Focal}} = -\alpha(1-p_t)^{\gamma}\log(p_t) \tag{4}$$

where:

- p_t is the predicted probability for the true class, i.e., $p_t = p$ for the positive class and $p_t = 1-p$ for the negative class.
- α is a weighting factor to adjust for class imbalance, typically set to 0.25.
- γ is the focusing parameter, commonly set to 2, which controls the down-weighting of well-classified examples.
- p is the predicted probability of the positive class.

The Focal Loss introduces a modulating factor $(1-p_t)^{\gamma}$ that reduces the relative loss for well-classified examples, thereby focusing training on hard, misclassified examples. It is particularly useful in tasks with class imbalance.

Discussion. Forensim leverages a combination of complementary loss functions to enhance accuracy and robustness in copy-move forgery detection (CMFD). Cross-Entropy Loss is employed for multi-class pixel classification, enabling the model to distinguish pristine, source, and target regions. InfoNCE Loss maximizes mutual information between similar patches by pulling positive pairs closer and pushing apart negative pairs, aiding in manipulation-aware representation learning. Dice Loss further improves segmentation quality by optimizing the overlap between predicted and ground-truth masks, while Focal Loss addresses class imbalance by emphasizing harder-to-classify pixels. The effectiveness of each component is demonstrated in the ablation study (Tab. 2), where the combination of Cross-Entropy and InfoNCE achieves the best trade-off across all metrics. Among them, Balanced Accuracy (BAcc) is particularly relevant in imbalanced settings, and is computed as the mean class-wise recall: $\text{BAcc} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{TP}_c}{\text{TP}_c+\text{FN}_c}$, where C is the number of classes and TP/FN denote true positives and false negatives for class c. Together, these loss functions enable Forensim to localize source and target regions more precisely, even in challenging complex forgery scenarios.

4. Additional Ablation Study on CMFD_Anything Dataset

To assess the impact of dataset composition on model performance, we conduct a dataset ablation study across different manipulation types and sources. As shown in Tab. 3, models trained exclusively on CASIA [4] or CoMoFoD [17] achieve limited performance, with maximum F1 scores of 31.6% and 29.2%, respectively. This highlights their limited diversity and domain coverage. Incorporating CMFD_Anything as a training source boosts generalization, while combining CASIA and CoMoFoD yields moderate improvements. The best performance is achieved when all datasets are combined, confirming that diverse training data significantly enhances the model's robustness across different manipulation types.

Table 3. Dataset ablation study. All models are evaluated on the CMFD_Anything test set with Forensim. We vary the training composition by dataset and manipulation type. Performance improves with diverse forgery exposure.

Training Data	MCC	F1 (Target)	AUC	BAcc
CASIA [4] only (Splicing)	0.418	0.316	0.598	0.683
CASIA [4] only (Copy-Move)	0.401	0.302	0.587	0.671
CASIA [4] only (Removal)	0.406	0.308	0.592	0.677
CoMoFoD [17] only (Copy-Move)	0.389	0.292	0.579	0.664
CASIA [4] + CoMoFoD [17]	0.510	0.374	0.623	0.712
CMFD_Anything only (Copy-Move)	0.622	0.590	0.682	0.766
All Combined for Forensim training	**0.681**	**0.624**	**0.700**	**0.812**

Method	CASIA v1				Columbia				DSO-1				NIST16				AVG			
	Fb	Wa	Wb	Wc	Fb	Wa	Wb	Wc	Fb	Wa	Wb	Wc	Fb	Wa	Wb	Wc	Fb	Wa	Wb	Wc
IF-OSN [19]	.513	.524	.507	.454	.741	.752	.756	.760	.484	.395	.416	.414	.315	.302	.292	.282	.513	.493	.493	.478
CAT-Net v2 [10]	.681	.508	.469	.206	**.964**	**.952**	**.958**	**.903**	.310	.247	.240	.237	.219	.238	.243	.244	.544	.486	.478	.398
MVSS-Net [2]	.469	.444	.480	.339	.752	.747	.758	.752	.356	.308	.354	.329	.305	.252	.300	.269	.471	.438	.473	.422
TruFor [7]	.716	.713	.676	.615	.797	.798	.835	.820	**.685**	**.465**	**.515**	**.469**	.338	.384	.308	.358	.634	.590	.584	.566
Forensim	**.729**	**.722**	**.685**	**.628**	.806	.806	.842	.828	.670	.452	.502	.455	**.346**	**.392**	**.315**	**.366**	**.638**	**.593**	**.586**	**.569**

Table 4. Pixel-level F1 (fixed threshold of 0.5) on images uploaded to social networks: Facebook (Fb), WhatsApp (Wa), Weibo (Wb), and WeChat (Wc). **Bold** = Best, Underline = Second-Best.

5. Additional Robustness Analysis on CoMoFoD Attacks

Figure 1 shows the number of correctly detected images across seven attack categories in the CoMoFoD dataset, where a prediction is considered correct if its pixel-level F1 score exceeds 30%. We compare several state-of-the-art copy-move forgery detection models, including BusterNet [20], MantraNet [21], DOA-GAN [8], TruFor [7], SparseViT [16], and our proposed Forensim. Forensim consistently outperforms all baselines across most attack categories—including challenging ones like Joint Compression (JC) and Noised Affine (NA)—demonstrating its strong robustness to various post-processing operations. Notably, SparseViT and TruFor also show strong performance, but Forensim achieves the highest number of correctly detected images overall, highlighting its effectiveness in real-world manipulation scenarios.

6. Additional Robustness Analysis on Social-Network Uploads.

As illustrated in Table 4, *Forensim* performs consistently better than other baselines across platforms (Fb/Wa/Wb/Wc) and datasets, indicating robustness to social-media degradations (pixel-level F1 at a fixed 0.5 threshold).

References

[1] Nimble challenge 2017 evaluation — nist. https://www.nist.gov/itl/iad/mig/nimble-challenge-2017-evaluation. (Accessed on 11/14/2020). 2, 3

[2] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14185–14193, 2021. 4

[3] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13, 2020. 3

[4] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, pages 422–426. IEEE, 2013. 2, 3, 4

[5] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12124–12134, 2022. 1

[6] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936. 2

[7] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Com-*

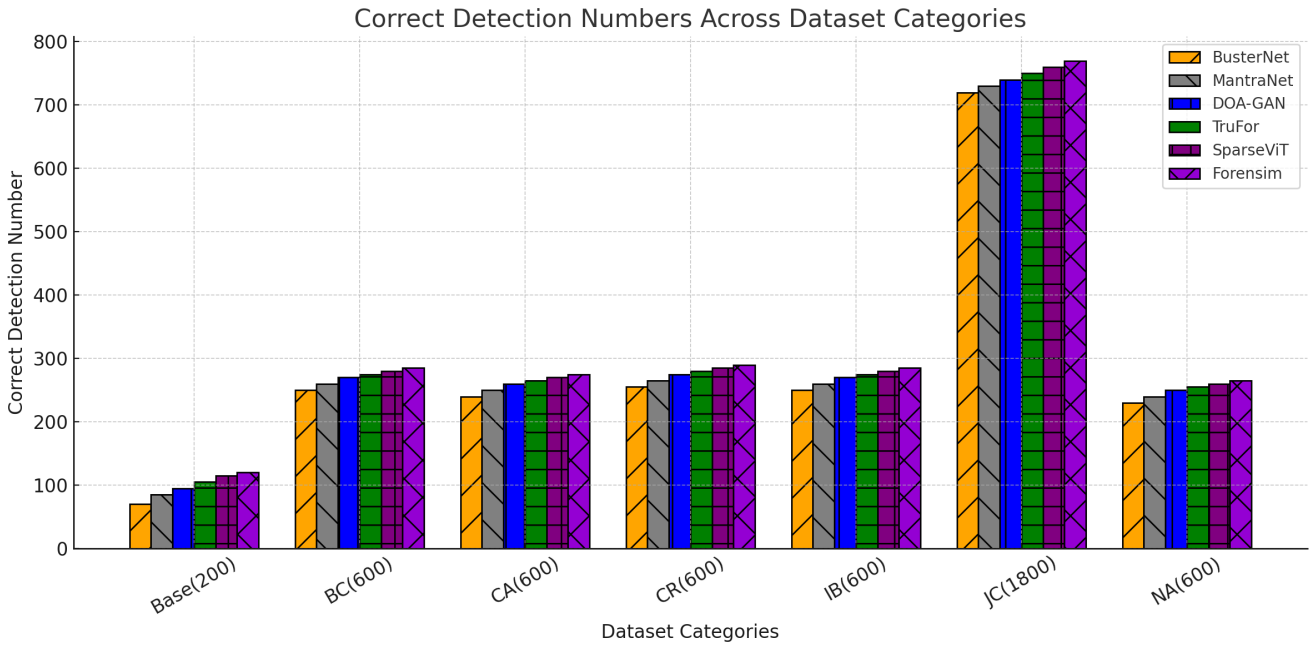

Figure 1. Comparison of correct detection numbers across different dataset categories for various forgery detection methods. The models compared include Adaptive-Seg, DenseField, BusterNet, DOA-GAN, and Forensim. DOA-GAN outperforms previous methods in most categories, while Forensim achieves the highest detection rate, slightly surpassing DOA-GAN. Each method is represented with a unique color and texture for clarity.

puter Vision and Pattern Recognition, pages 20606–20615, 2023. 4

[8] Ashraful Islam, Chengjiang Long, Arslan Basharat, and Anthony Hoogs. Doa-gan: Dual-order attentive generative adversarial network for image copy-move forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4676–4685, 2020. 1, 4

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 3

[10] Myung-Joon Kwon, In-Jae Yu, Seung-Hun Nam, and Heung-Kyu Lee. Cat-net: Compression artifact tracing network for detection and localization of image splicing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 375–384, 2021. 4

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[12] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[13] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 2

[14] Tian-Tsong Ng, Jessie Hsu, and Shih-Fu Chang. Columbia image splicing detection evaluation dataset. *DVMM lab. Columbia Univ CalPhotos Digit Libr*, 2009. 2, 3

[15] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[16] Lei Su, Xiaochen Ma, Xuekang Zhu, Chaoqun Niu, Zeyu Lei, and Ji-Zhe Zhou. Can we get rid of handcrafted feature extractors? sparsevit: Nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7024–7032, 2025. 4

[17] Dijana Tralic, Ivan Zupancic, Sonja Grgic, and Mislav Grgic. Comofod—new database for copy-move forgery detection. In *Proceedings ELMAR-2013*, pages 49–54. IEEE, 2013. 2, 3, 4

[18] Bihan Wen, Ye Zhu, Ramanathan Subramanian, Tian-Tsong Ng, Xuanjing Shen, and Stefan Winkler. Coverage—a novel database for copy-move forgery detection. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 161–165. IEEE, 2016. 2, 3

[19] Haiwei Wu, Jiantao Zhou, Jinyu Tian, Jun Liu, and Yu Qiao. Robust image forgery detection against transmission over online social networks. *IEEE Transactions on Information Forensics and Security*, 17:443–456, 2022. 4

[20] Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Busternet: Detecting copy-move image forgery with source/target localization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 168–184, 2018. 2, 3, 4

[21] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9543–9552, 2019. 1, 4

[22] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 3