Towers of Babel: Combining Images, Language, and 3D Geometry for Learning Multimodal Vision — Supplementary Material —

Contents

1. Dataset Visualizations and Details **1**

2. Implementation Details **1**

3. Ablation Study **3**

4. Additional Classification Results **3**

5. Additional Qualitative Results **5**

1. Dataset Visualizations and Details

We show captions with spatial connectors and their corresponding images in Figure 1 to illustrate the richness of part interactions contained within our dataset. Please refer to the accompanying `samples.html` for samples from our dataset that are used for training. These are grouped according their concepts.

Data distributions. Figure 2 shows the distribution of captions by the number of words. Figure 3 shows the number of data samples by landmark identity sorted by size. Figure 4 shows the number of captions in the top 10 languages. The caption's language is detected according to [8].

2. Implementation Details

2.1. Dataset construction

We use COLMAP [11] version 3.6 for building 3D reconstructions. The SIFT [6] peak threshold is set to 0.03. To find image matches, we use vocabulary tree matching [12] using the pretrained vocabulary tree with 1M visual words. For landmarks that have reconstructions in the MegaDepth dataset [5] (we have 44 shared landmarks), external images from their dataset were added to assist reconstruction. Original high resolution images are used for reconstruction. However, for training purposes, we use resized images with the shorter dimension set to 200 pixels. We also release a higher resolution version in our dataset, where the longer dimension is set to 1200 pixels.

2.2. Network architecture

Figure 5 shows the structure of our network. The structure closely follows the network proposed in Araslanov *et al*. [1]. For completeness, we briefly summarize the network architecture here. We use a Resnet-50 backbone to extract both low-level and high-level features (which is pretrained on ImageNet). Atrous Spatial Pyramid Pooling (ASPP) [2] augments the ResNet features by gathering information at different scales. A Global Cue Injection (GCI) module [1] infuses global cues from deep layers into low-level features derived from the shallow layers of ResNet. The stochastic gate [1] aims to mitigate overfitting introduced by errors in the pseudo-ground truth used during training. The 3D consistency loss is computed on the features before unnormalized score maps are computed. The classification score y is computed according to [1], summing a normalized Global Weighted Pooling (nGWP) term and a focal penalty term (Equation 3 and Equation 5 in their paper).

2.3. Training details

Our models are implemented in PyTorch [9]. We train our model using the Adam optimizer [4] with weight decay 5×10^{-4}, and using default Adam parameters. The model is trained for 25 epochs with learning rate decay occurring at the 15^{th} and 20^{th} epoch. Following [1], we pretrain the model without $\mathcal{L}_{cls}^{pix}$ for 5 epochs. In all our experiments, learning rate decay is performed using a factor of 0.1. For all experiments with $\mathcal{L}_{\text{3D}}$, the balancing coefficient is set to 0.3 (i.e., the 3D constrastive loss is multiplied by this coefficient). The temperature used in the 3D constrastive loss is set to the default value of 0.07 and the number of negatives is 16. All models are pretrained on ImageNet.

As the images in WikiScenes are of varying resolution, we perform a random resized crop operation to convert each image to $[224 \times 224]$ training samples. The scale factor of the random resized crop is sampled from the range $[0.9, 1.0]$. Random horizontal flipping and color jittering are also performed to augment the data. The brightness, contrast, saturation and hue parameters are set to $[0.3, 0.3, 0.3, 0.1]$, respectively, in the color jittering step. We balance the size

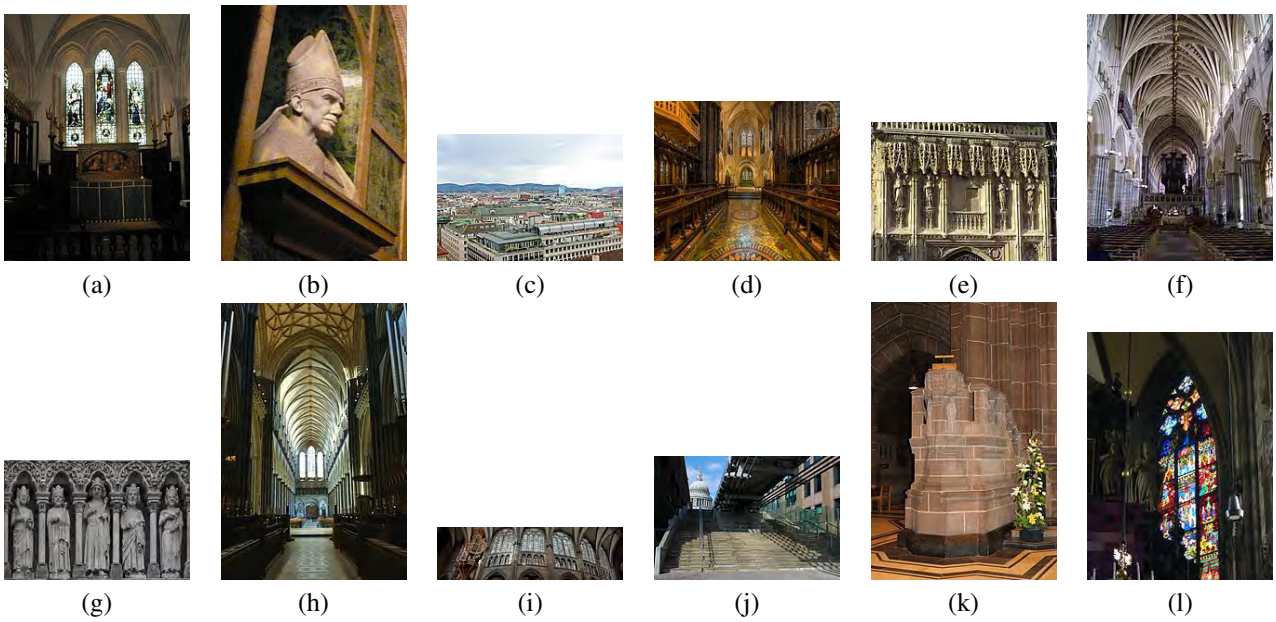


Figure 1: **Example images from WikiScenes.** Corresponding captions are: (a) *Altar behind the main quire at Southwark Cathedral.* (b) *Bishop Gregorio Modrego over his tomb in cathedral of Barcelona, by Fredric Marès.* (c) *Went to the top of the bell tower to see the views looking over the city of Vienna.* (d) *The choir of Christ Church Cathedral in Dublin, Ireland, looking east towards the sanctuary.* (e) *Statues above the main entrance of Canterbury Cathedral: (left to right) Augustine of Canterbury, Lanfranc, Anselm of Canterbury and Thomas Cranmer.* (f) *The nave of Exeter Cathedral From the west end of the nave looking towards the crossing with its 17th century organ.* (g) *Amiens, France: Fassade detail of the Cathedrale of Amiens, showing the right group of sculptures under the rosette window.* (h) *Salisbury Cathedral Looking towards the West Front, from the Quire.* (i) *The Silbermann organ in Strasbourg cathedral, view from below with the nave windows.* (j) *The Dome of St Paul's Cathedral viewed from the river bank below the Millennium Bridge.* (k) *Sandstone pulpit next to the north transept of Liverpool Anglican Cathedral.* (l) *Window with medieval glass painting behind the high altar in St. Stephen's Cathedral, Vienna.*

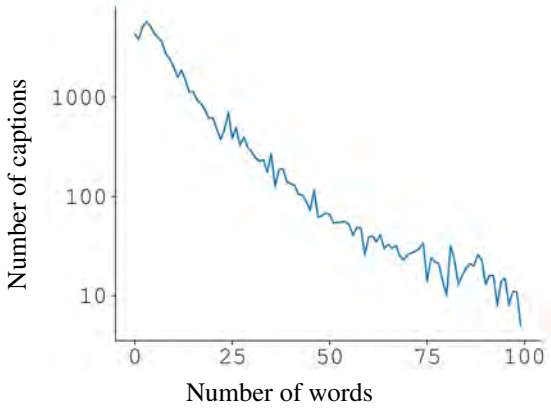


Figure 2: Distribution of image captions by number of words (y-axis is plotted on a log scale).

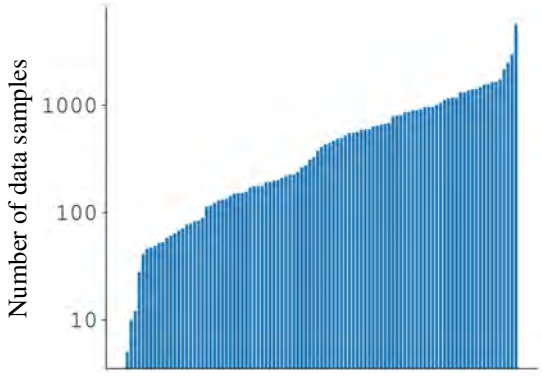


Figure 3: Number of images paired with textual descriptions by landmarks (sorted). The y-axis is plotted on a log scale.

of the different classes by resampling. The balanced dataset contains roughly 900 images in each class.

2.4. Additional 2D segmentation details

Our model predicts both image-level classification scores (*i.e.*, y, see Figure 5) and pixel-wise normalized segmentation score maps (*i.e.*, y_{pix}, see Figure 5) that also include a background score, in addition to scores for each of the C semantic concepts. Following [1], the background score is

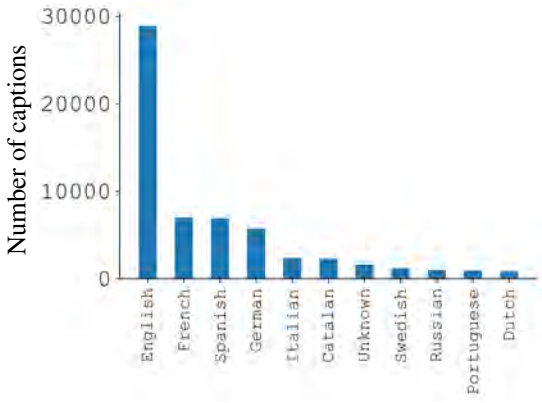


Figure 4: Number of captions in the top 10 languages. "Unknown" denotes captions that are not recognizable, such as date, URL or null strings.

.

weakened by a power function (set empirically to the 4^{th} power in our experiments). We first take the maximal value in y to select the image-level label (we only consider images that contain a single label). Then, the 2D segmentation mask is comprised of all pixels whose score corresponding to the selected image-level concept surpasses that of the (weakened) background score.

2.5. Additional 3D segmentation details

For each point in a 3D model, we first gather classification scores from its 2D projection in different views. The scores are averaged before applying softmax function to obtain the classification score of the 3D point. Points with scores higher than a predetermined threshold φ are considered foreground, and the remaining points are considered ambiguous and are therefore not rendered in our 3D visualizations. For quantitative evaluation, we provide results for $\varphi = 0.5$ and $\varphi = 0.75$. Our visualizations are rendered using a threshold of $\varphi = 0.5$.

2.6. Caption-based image retrieval experiment

To perform the caption-based retrieval experiment, we run the command for fine-tuning from the multi-task trained model[1]. We define two new tasks, one for the baseline model (that uses only original image-caption pairs) and another for the 3D-augmented model (that also uses 3D-augmented pairs). Both models are trained for 12 epochs, using their (unmodified) configurations.

Regarding evaluation, to compute our proxy semantic measure S, we followed their retrieval evaluation and construct a batch of 1000 images from validation, however, in our case, this batch includes all labeled images from unseen images (774 images in total) and additional randomly-selected unlabeled images. We use these labels for evaluating whether or not the label of the retrieved images agree with the label of the target image.

[1]Available here: `https://github.com/facebookresearch/vilbert-multi-task`

3. Ablation Study

We perform an ablation study to analyze our design choices for the 3D-consistency regularization. We replace our 3D contrastive loss with the following alternatives:

3D MSE loss. We compute a simple MSE loss between the features of corresponding pixels:

$$\mathcal{L}_{\text{3D}}^{\text{MSE}} = \|F(p) - F(p^+)\|_2^2.$$

3D Triplet loss. We select one negative pixel p^- and compute the following 3D loss:

$$\mathcal{L}_{\text{3D}}^{\text{triplet}} = \max(0, \|F(p)-F(p^-)\|_2^2 - \|F(p)-F(p^+)\|_2^2 + m),$$

where m is a margin value (set empirically to 3).

3D intra-image contrastive loss. In the main paper, we introduce a 3D contrastive loss $\mathcal{L}_{\text{3D}}(\textit{inter-image sampling})$, where the negative pixels $\{p_i^-\}$ are sampled from other images in the batch. We change the sampling strategy such that all the negative pixels are selected from other regions in the same image to obtain $\mathcal{L}_{\text{3D}}(\textit{intra-image sampling})$. Specifically, the points p_i^- are sampled uniformly in I_2, outside a box of size $[w\ /\ 4, h\ /\ 4]$ around p^+.

Results are reported in Table 1. As illustrated in the table, we can improve classification performance using a variety of loss configurations. Our 3D contrastive loss, using both inter-image and intra-image sampling strategies, yield the most significant improvements.

Following prior work [1], our semantic classification loss is composed of two terms, where $\mathcal{L}_{\text{pix}}^{\text{cls}}$ is a self-supervised loss over pixelwise predictions that is applied starting at the 6-th epoch. Classification performance is roughly the same when this self-supervised loss is not used. Specifically, mAP increases from 52.0 to 53.8 for WS-U and decreases from 75.3 to 73.4 for WS-K. The gaps to the baseline model mostly remain unchanged (3.3% improvement for WS-U and 1.8% improvement for WS-K).

4. Additional Classification Results

Figure 6 shows a confusion matrix for our image classification model. We observe that many of the mistakes are understandable, given the hierarchical nature of our data. For example, both "tower" and "portal" are part of a "facade", and an "altar" is often placed inside a "chapel".

To further explore the hierarchical structure of semantics in our dataset, we associate images with *ancestor* labels by

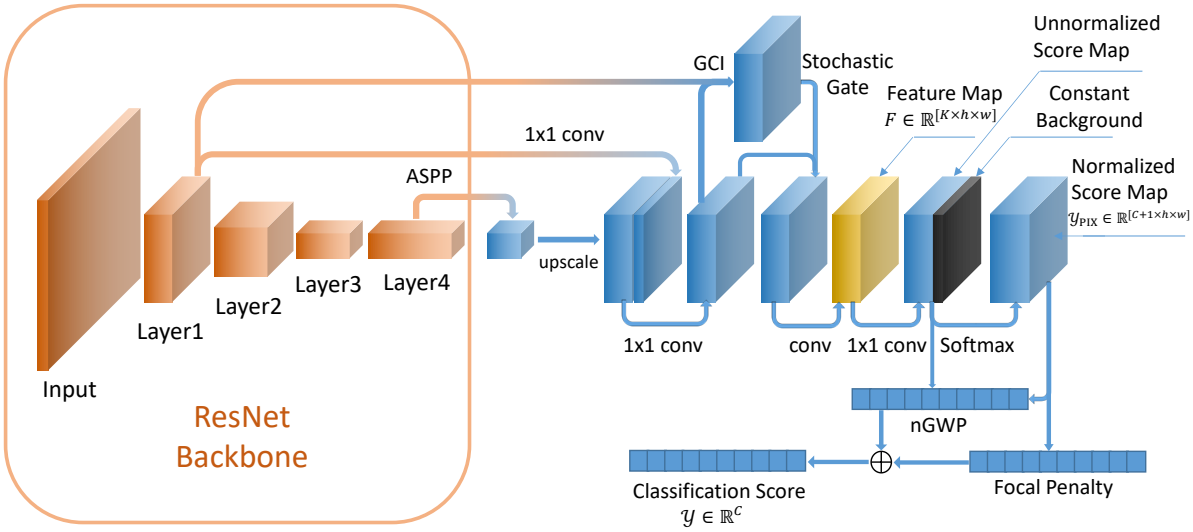


Figure 5: Our classification network architecture.

Test Set	Model	mAP	facade	window	chapel	organ	nave	tower	choir	portal	altar	statue
WS-K	Baseline (w/o 3D loss)	70.8	87.2	89.2	60.2	89.7	85.8	64.1	61.5	68.0	50.0	52.0
	w/ $\mathcal{L}_{3D}^{MSE}$	71.4	86.4	88.3	53.1	89.4	86.1	65.7	62.0	69.7	52.5	60.3
	w/ $\mathcal{L}_{3D}^{triplet}$	72.1	88.5	**90.5**	55.6	86.0	**86.4**	**66.5**	65.0	68.4	50.2	**63.4**
	w/ $\mathcal{L}_{3D}$(*intra-image sampling*)	73.3	**90.4**	87.1	62.9	90.3	85.8	62.1	75.9	68.4	52.8	57.1
	w/ $\mathcal{L}_{3D}$(*inter-image sampling*)	**75.3**	90.0	88.5	**68.7**	**90.7**	85.7	61.1	**77.2**	**76.5**	**54.4**	59.9
WS-U	Baseline (w/o 3D loss)	48.3	71.0	92.2	10.7	57.3	71.0	**53.4**	43.6	31.1	25.8	27.1
	w/ $\mathcal{L}_{3D}^{MSE}$	49.5	70.6	94.3	10.9	61.8	73.7	50.8	40.9	41.3	21.6	28.9
	w/ $\mathcal{L}_{3D}^{triplet}$	49.9	73.1	**94.9**	9.9	53.7	74.7	47.5	40.8	29.1	39.4	35.6
	w/ $\mathcal{L}_{3D}$(*intra-image sampling*)	**52.5**	75.8	94.1	**16.7**	**62.5**	75.4	50.4	**44.5**	**43.0**	24.4	38.4
	w/ $\mathcal{L}_{3D}$(*inter-image sampling*)	52.0	**77.7**	93.4	16.5	49.4	**77.3**	46.1	44.1	35.2	**39.9**	**40.0**

Table 1: Classification performance using different types of 3D-consistency regularizations. We report mean average precision (mAP), and per distilled–concept average precision (AP). Please refer to Section 3 for more details on the different configurations. Performance is reported on images from two different tests sets corresponding to known landmarks (WS-K) and unseen landmarks (WS-U). The best result for each test set and column are highlighted in bold.

considering the concepts present in its hierarchy of WikiCategories. Unlike prior works that require manually annotating such hierarchical labels (e.g., [7]), we obtain these automatically, leveraging the hierarchical structure of Wikimedia Commons. In Figure 7, we visualize these hierarchical relationships. Many of these relationships can also be observed from the confusion matrix of our model in Figure 6. We also observe additional intuitive connections such as an image associated with "window" also being associated with larger structures such as "facade" and "nave"; a "statue" can be placed on various structures, and so on.

Finally, to further validate the effectiveness of our 3D loss, we take off-the-shelf networks dedicated for classifi-

	Resnet-50 [3]		MobileNetV2 [10]	
Test Set	w/o $\mathcal{L}_{3D}$	w/ $\mathcal{L}_{3D}$	w/o $\mathcal{L}_{3D}$	w/ $\mathcal{L}_{3D}$
WS-K	68.5	**73.9**	77.1	**79.6**
WS-U	48.7	**52.3**	50.2	**53.4**

Table 2: Evaluating the effectiveness of our 3D contrastive loss on off-the-shelf classification models. For each model, we report mAP. The best results are highlighted in bold.

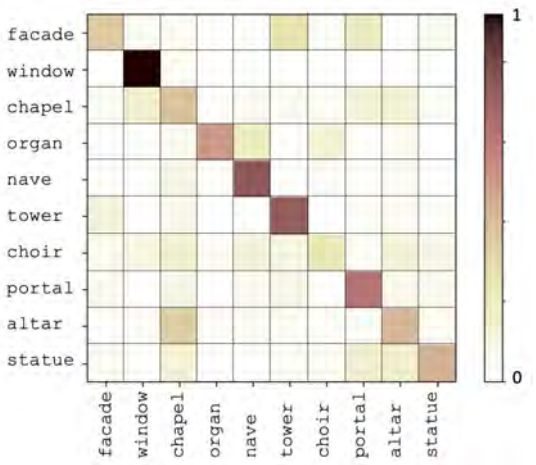


Figure 6: Confusion matrix of our classification model on unseen landmarks (the WS-U test set). Ground-truth concept labels correspond to rows, and predicted concept labels to columns. Each row is normalized such that a cell indicates the probability of a classification given the ground-truth label.

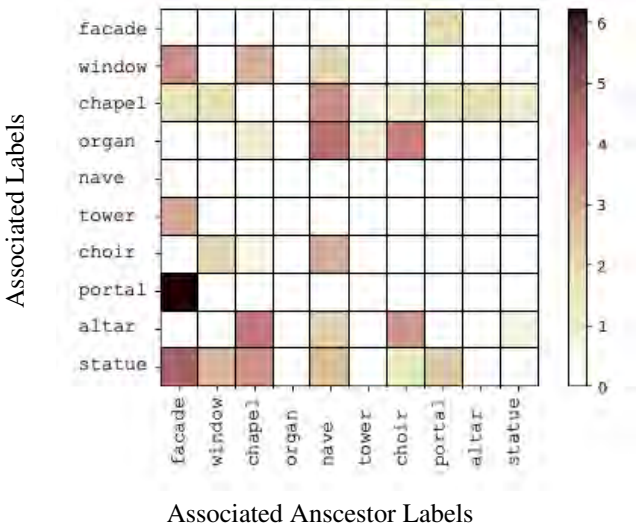


Figure 7: Associating images with labels and anscestor labels. Above we visualize (in log scale) the co-occurrence of concepts as labels and ancestor labels.

cation and repeat the experiment of testing classification performance with and without our 3D contrastive loss. For this experiment, all models are trained for 10 epochs with a learning rate decay at the 6^{th} epoch. Both Resnet-50 and MobileNetV2 are pretrained on ImageNet.

Results are reported in Table 2. As illustrated in the table, our 3D contrastive loss consistently boosts classification performance, even for off-the-shelf models.

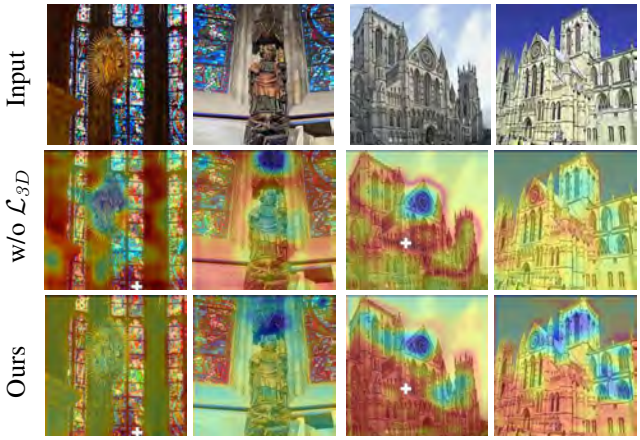


Figure 8: Visualizing distances in feature space for unseen landmarks. For each image pair, we select a random pixel in the left image (marked in white) and visualize the distance to all other pixels from the selected pixel (marked in white) with and without our 3D contrastive loss. Warmer colors correspond to smaller distances. As illustrated above, distances in feature space are more semantically meaningful on the model trained with the 3D contrastive loss (see, for instance, distances on the windows in the left pair). Our model is also more robust against large motion and appearance variations between the images (as illustrated on the right).

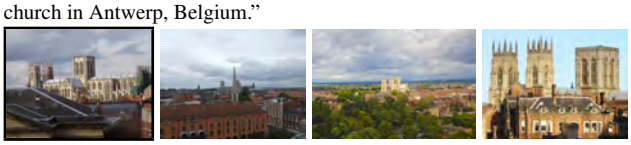
5. Additional Qualitative Results

We show additional image segmentation results on test images from the WS-K test set in Figure 10 and Figure 11. As illustrated in the figures, the model is more successful with segmenting certain concepts, such as “tower”, “portal” or “window”. Some concepts, such as “chapel” yield noisier segmentation results. We show 3D segmentation results for landmarks in WS-K in Figure 12.

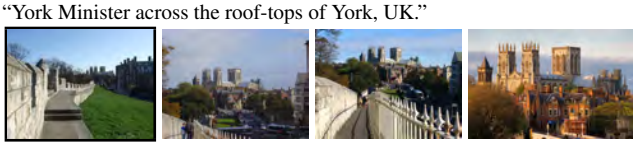
We visualize the learned features for two image pairs in Figure 8. As the figure illustrates, distances in feature space are more semantically meaningful on the model trained with the 3D contrastive loss. For example, only pixels on the windows yield small distances using our model (left image pair). Our model is also more robust against large motion and appearance variations between the images.

We show additional caption-based image retrieval results in Figure 9, mostly for images not-labeled with one of the semantic concepts we compute according to the method described in the main paper. As demonstrated in the figure, the model can also align more generic semantic concepts to our images. However, as we perform 3D-augmentations, the model is less aware of appearance-based differences. For example, see the bottom row in the figure, where the retrieved images are not captured “at night”.

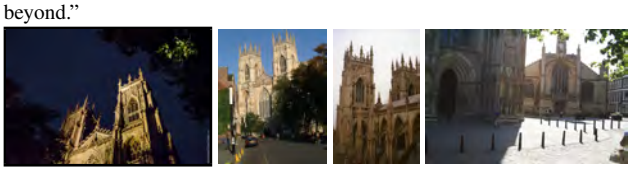
"Neogothic portal of Our Lady's Cathedral, Antwerp, by Jean Baptist van Wint (1829-1906). The Cathedral of Our Lady is a Roman Catholic parish church in Antwerp, Belgium."



"York Minister across the roof-tops of York, UK."



"York city walls pathway Looking towards Lendal Bridge and the Minster beyond."



"York Minster at night (2012)"

Figure 9: Retrieving images from captions of The Cathedral of Our Lady and York Minister (landmarks not seen during training). Above we show the top three retrievals next to the reference image (left with black border) that corresponds to the query caption beneath. Note that this query image is not seen by the network—just the caption—and so we only show this image for reference. In the bottom row, we demonstrate that our model is less sensitive to appearance-based descriptions—in this case, the retrieved images are not captured "at night". This can be attributed to our 3D augmentations, which are unaware of appearance changes (thus allowing to focus on part-based *scene* semantics instead).

References

[1] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *CVPR*, 2020.

[2] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *PAMI*, 40(4):834–848, 2018.

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

[4] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[5] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018.

[6] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004.

[7] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *CVPR*, 2019.

[8] Shuyo Nakatani. Language detection library for java, 2010.

[9] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[10] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018.

[11] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016.

[12] Johannes Lutz Schönberger, True Price, Torsten Sattler, Jan-Michael Frahm, and Marc Pollefeys. A vote-and-verify strategy for fast spatial verification in image retrieval. In *Asian Conference on Computer Vision (ACCV)*, 2016.

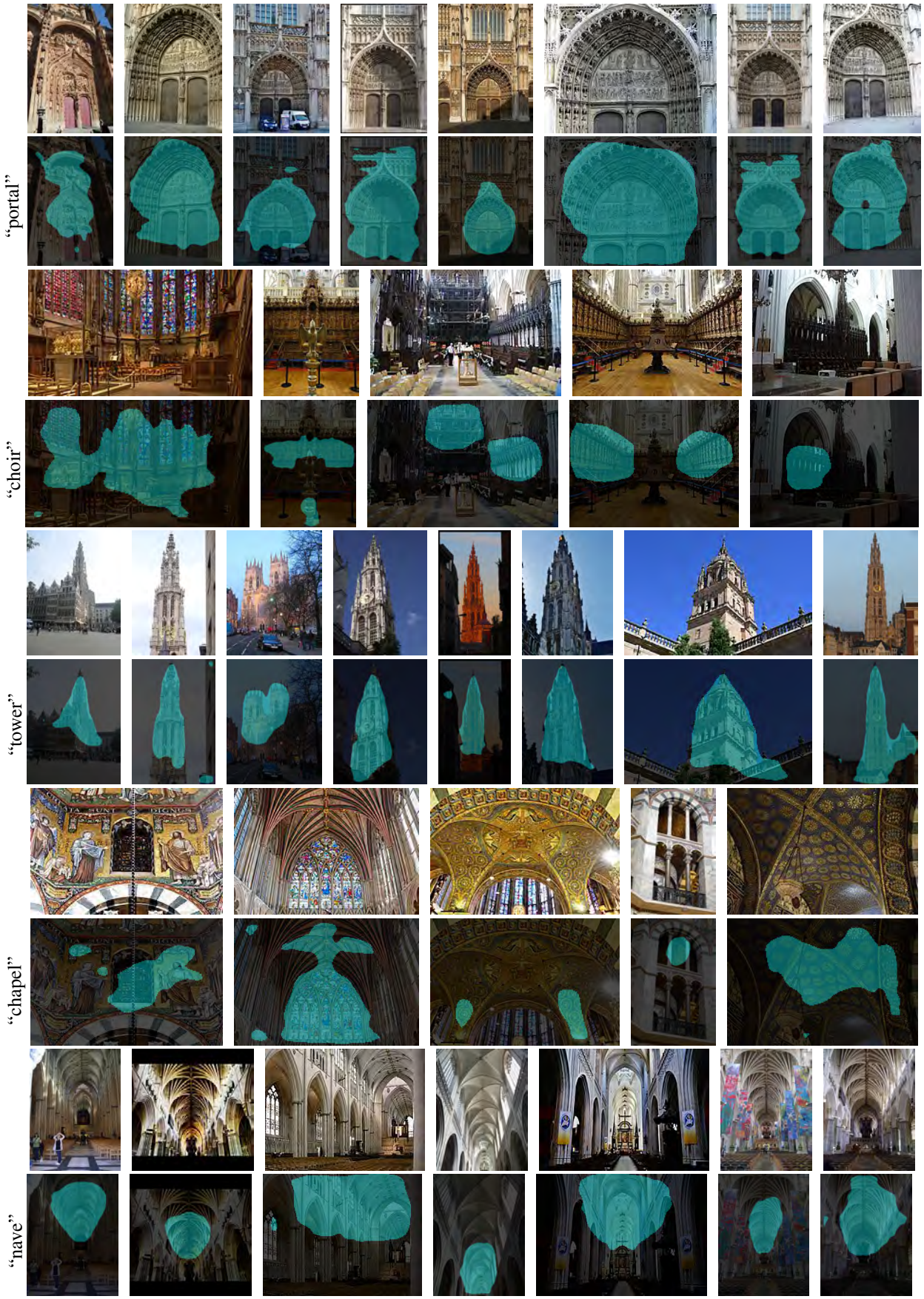


Figure 10: **2D Segmentations on correctly classified unseen images,** segmented as “portal”, “choir”, “tower”, “chapel” and “nave”.

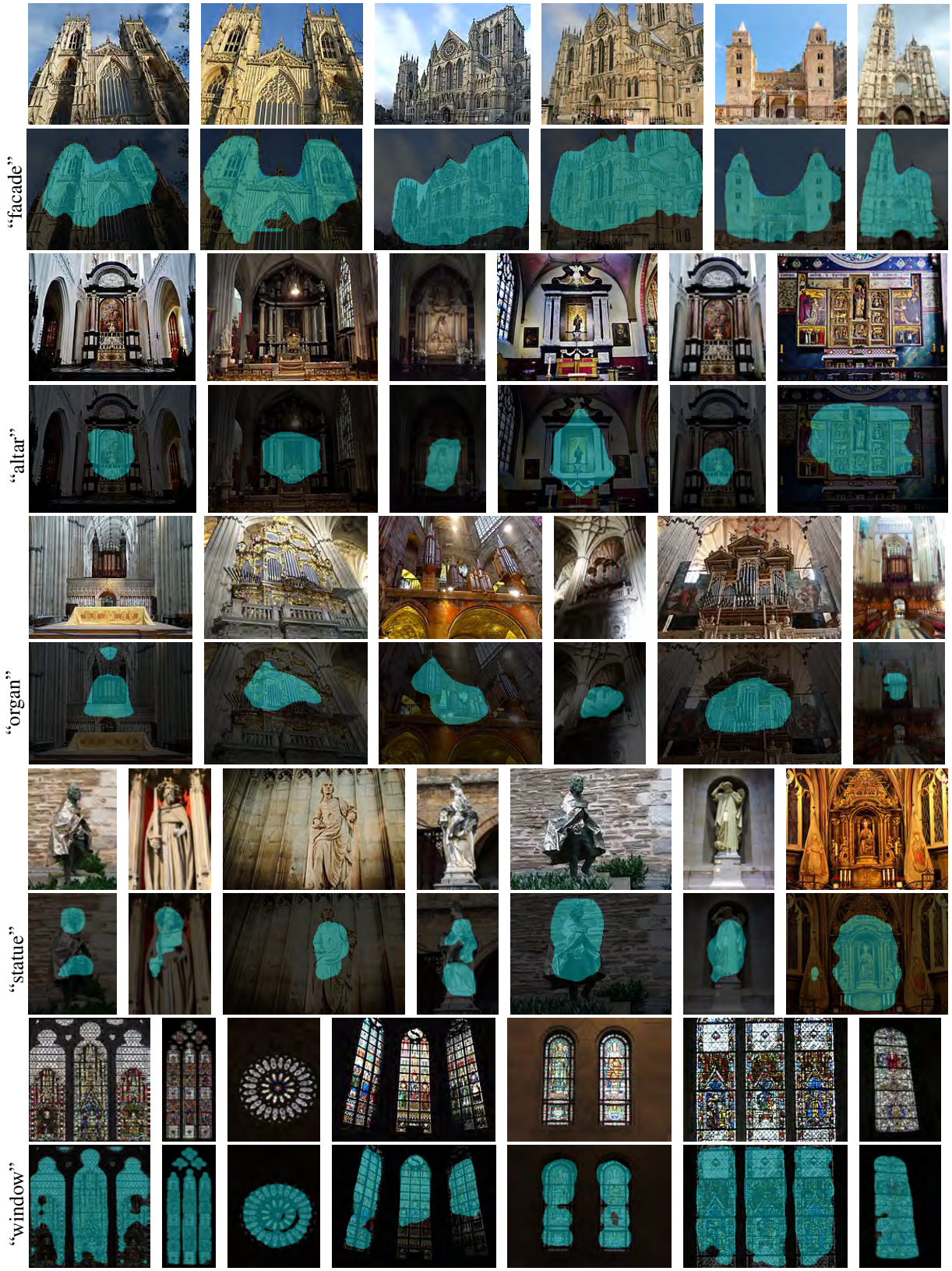


Figure 11: **2D Segmentations on correctly classified unseen images.** Hightlighted pixels are segmented as “facade”, “altar”, “organ”, “statue” and “window”.

Notre-Dame de Strasbourg (exterior)

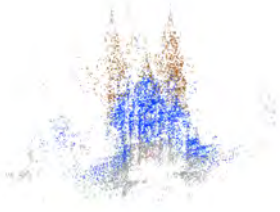
Cathedral of Barcelona (exterior)

Notre-Dame de Reims (exterior)

Saint Isaac's Cathedral (exterior)

St. Stephen's Cathedral (exterior)

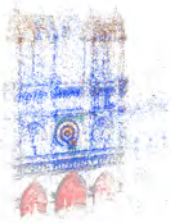
Amiens Cathedral (exterior)



Santiago de Compostela Cathedral (exterior)

St Paul's Cathedral (exterior)

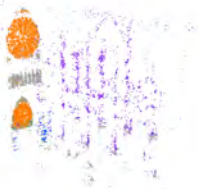
Notre-Dame de Paris (exterior 1)



Notre-Dame de Paris (exterior 2)

Ulm Minster (exterior)

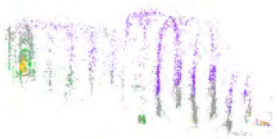
Duomo di Milano (interior)



Notre-Dame de Reims (interior)

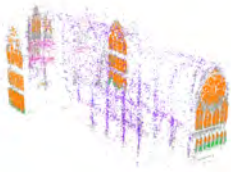
Cathédrale Saint-André de Bordeaux (interior)

Saint Isaac's Cathedral (interior)

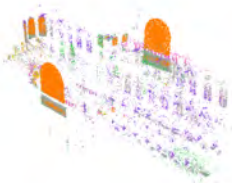


St. Stephen's Cathedral (interior)

Amiens Cathedral (interior)



Metz Cathedral (interior)



Notre-Dame de Paris (interior)

León Cathedral (interior)

Figure 12: Segmenting 3D reconstructions. Above we show segmentation results for landmarks seen during training. 3D points not associated with concepts are colored in gray. Color legend of segmented points: *nave*, *chapel*, *organ*, *altar*, *choir*, *statue*, *portal*, *facade*, *tower*, *window*.