DeepChoice: Learning View Weighting for Image-Guided 3D Semantic Segmentation

Antoine Carreaud[1,2], Digre Frinde[2], Shanci Li[2], Jan Skaloud[1], Adrien Gressin[2]

[1] ESO lab. EPFL, 1015 Lausanne, Switzerland - (firstname.lastname)@epfl.ch
[2] University of Applied Sciences Western Switzerland (HES-SO / HEIG-VD),
Yverdon-les-Bains, Switzerland - (firstname.lastname)@heig-vd.ch

Keywords: 3D semantic segmentation, multi-view fusion, image-to-point projection, LiDAR, photogrammetry.

Abstract

Multi-view image-to-point label transfer is an effective strategy for 3D semantic segmentation, but its performance largely depends on how predictions from multiple image observations are fused for each 3D point. Most existing pipelines rely on hard voting or handcrafted weighting rules, which do not explicitly learn the reliability of each view under varying geometric and image-quality conditions. In this paper, we introduce **DeepChoice**, a lightweight view-weighting module for image-guided 3D semantic segmentation. For each visible observation of a 3D point, DeepChoice exploits a compact set of visibility cues, including incidence angle, range, contrast, sharpness, signal-to-noise ratio, and saturation, to predict normalized per-view weights used to aggregate 2D semantic class probabilities into final 3D point-wise predictions. The method is sensor-agnostic, requires no meshing, and can be integrated as a drop-in replacement for standard multi-view fusion rules. Experiments on the full *GridNet-HD* benchmark show that DeepChoice improves over hard voting by **3.85** mIoU points and over mean-probability fusion by **1.26** mIoU points, while reducing the gap with the *AnyView* oracle upper bound. The largest gains are observed on thin and difficult classes such as conductors, pylons, and insulators. Furthermore, a complementary evaluation on the *Images&PointClouds Cultural Heritage* dataset shows that the proposed weighting strategy remains beneficial under a very different acquisition context and scene structure, yielding a **1.55** mIoU point improvement over hard voting and a **0.48** mIoU point improvement over mean-probability fusion. A compact Transformer variant provides the best trade-off between accuracy and model size, outperforming a larger MLP-based alternative. These results show that learning how to weight views is a simple yet effective way to strengthen image-guided 3D semantic segmentation pipelines. Code is publicly available at `https://huggingface.co/heig-vd-geo/DeepChoice`.

1. Introduction

Accurate 3D semantic segmentation is a key component of modern scene understanding pipelines, with applications ranging from digital twin generation and infrastructure monitoring to cultural-heritage digitization and autonomous inspection. In many practical settings, high-quality semantic information is first predicted in images and then transferred to a 3D model such as a point cloud or a mesh through calibrated multi-view projection (Hermans et al., 2014, McCormac et al., 2017, Peters et al., 2023). This image-guided strategy is attractive because it leverages the maturity of 2D semantic segmentation in terms of model and dataset availability, while avoiding the need to train a dedicated 3D network.

A central challenge in such pipelines is the fusion of multiple image observations associated with the same 3D point. In practice, a point may be visible in several images captured from different distances, under different viewing angles, and with different local image quality. Yet most existing methods still rely on simple fusion rules such as hard majority voting or handcrafted weighting schemes based on geometric heuristics (Pellis et al., 2025a, Huang et al., 2025, Carreaud et al., 2025). These strategies remain effective baselines, but they do not explicitly learn which observations are the most reliable for semantic transfer.

In this work, we address this limitation with **DeepChoice**, a lightweight module that learns to weight image observations before fusing them into point-wise 3D semantic predictions. For each visible point-view pair, we compute a compact observation descriptor built from geometric and image-quality cues, namely incidence angle, range, local contrast, sharpness, local signal-to-noise ratio, and saturation. DeepChoice then predicts normalized per-view weights that are used to aggregate the corresponding per-class 2D probability vectors into a final semantic distribution for each 3D point. This method is entirely post-hoc with regard to the upstream image segmenter. It requires no meshing and can be incorporated into existing image-to-point fusion pipelines with only minor adjustments.

The proposed approach is motivated by a simple observation: not all views of a 3D point are equally informative. A point observed from far away, under a grazing angle, or in a locally blurred image region should not influence the final 3D prediction as much as a well-centered observation. Rather than relying on fixed handcrafted rules, we let the model learn this weighting policy directly from point-level supervision. In this sense, DeepChoice does not relearn semantics from raw image content; instead, it learns how to exploit already available 2D semantic evidence more effectively during 3D fusion. Our formulation is also related in spirit to learned view-aggregation strategies explored in image-LiDAR fusion (Robert et al., 2022), although we focus here on post-hoc semantic fusion in image-guided 2D-to-3D transfer pipelines.

We evaluate DeepChoice on the full *GridNet-HD* benchmark (Carreaud et al., 2026a) and further provide a complementary evaluation on the *Images&PointClouds Cultural Heritage* dataset (Pellis et al., 2025b). Our experiments show that learned view weighting consistently improves over standard non-learned fusion strategies, with particularly clear gains on thin and difficult classes such as conductors and insulators. We also compare two lightweight weighting architectures, an MLP

and a compact Transformer, and show that the Transformer provides the strongest performance-to-parameter ratio. Beyond the quantitative gains, these results indicate that view fusion itself is an important and underexplored source of improvement in image-guided 3D semantic segmentation.

Our main contributions are as follows:

- We introduce **DeepChoice**, a lightweight and sensor-agnostic learned view-weighting module for image-guided 3D semantic segmentation.
- We formulate multi-view 2D-to-3D fusion as a visibility-aware probability aggregation problem, using compact geometric and image-quality descriptors for each point-view observation.
- We demonstrate on the full *GridNet-HD* benchmark that learned view weighting improves over hard voting and non-learned score fusion, while remaining computationally lightweight.
- We provide a complementary evaluation on the *Images&PointClouds Cultural Heritage* dataset, showing that the proposed fusion strategy remains beneficial in a markedly different setting.

The remainder of the paper is organized as follows. Section 2 reviews related work on multi-view image-guided 3D semantic segmentation and fusion strategies. Section 3 presents the proposed DeepChoice formulation. Section 4 reports the experimental setup, main results, and ablation studies.

2. Related Work

We focus on image-guided 3D semantic segmentation pipelines that follow the same high-level decomposition as our method: semantic predictions are first produced in 2D, then transferred to a 3D support, and finally fused across multiple image observations. This family includes methods that aggregate semantic evidence on 3D points, surfels, volumetric grids, or meshes, but the central methodological choice remains the same: **how should multiple views be combined once they have been associated with the same 3D element?** A useful distinction is whether fusion operates on *hard labels*, where each view contributes only its winning class (Pellis et al., 2022, Pellis et al., 2025a, Carreaud et al., 2022), or on *confidence-preserving scores*, where the full per-view class distribution is retained until after 3D aggregation (Hermans et al., 2014, McCormac et al., 2017, Peters et al., 2023, Kundu et al., 2020, Mascaro et al., 2021).

Hard-label fusion. The simplest strategy is to project the predicted class of each image into 3D space and assign the most frequently occurring class. This majority-vote logic is attractive due to its simplicity and ease of deployment. For example, it is used, in the "mostly voted class" strategy of (Pellis et al., 2022, Pellis et al., 2025a) and as a naive baseline in the multi-view label-transfer pipeline of (Peters et al., 2023). A similar classify-then-vote baseline is also reported in OpenScene, where 3D points are assigned labels by majority voting over multi-view image predictions (Peng et al., 2023). Related application-driven pipelines also rely on hard image-to-3D transfer without explicitly preserving per-view confidence information during fusion (Carreaud et al., 2022). While effective as a baseline, hard-label fusion treats weak and strong observations equally once projected to 3D.

Weighted hard fusion and confidence-aware transfer. Several works retain the discrete label-selection logic while modulating the contribution of each observation. For example, image labels can be transferred to photogrammetric 3D data, with conflicts resolved using a weighted label-selection rule in which the highest-weighted label is retained (Stathopoulou and Remondino, 2019). In a different setting, 2D3DNet uses multi-view 2D predictions to derive pseudo-labels for 3D supervision, together with confidence-aware filtering to reduce noisy transferred supervision (Genova et al., 2021). These methods already acknowledge that not all views should contribute equally, but they do not formulate the problem as learning a continuous per-view fusion weight for post-hoc 2D-to-3D semantic score aggregation.

Score-preserving fusion and structured refinement. A second line of work keeps the full class-confidence vector during aggregation. One representative approach transfers 2D class probabilities to 3D points, aggregates them across observations in a Bayesian-style formulation, and refines the resulting 3D labels with a CRF (Conditional Random Field) regularizer (Hermans et al., 2014). SemanticFusion follows the same general principle on surfel maps, maintaining and recursively updating per-class surfel distributions as new views are observed (McCormac et al., 2017). Another approach relies on adaptive volumetric semantic fusion, followed by CRF-based refinement on reconstructed 3D surfaces (Jeon et al., 2018). Related multi-view methods have also explored score-preserving fusion in different forms. Virtual Multi-view Fusion revisits multi-view semantic segmentation of 3D meshes by rendering multiple virtual views and fusing their predictions on mesh vertices (Kundu et al., 2020). Diffuser formulates the 2D-to-3D transfer stage as a graph-based label diffusion problem, combining multi-view image predictions and 3D geometric relationships to produce more spatially consistent 3D segmentations (Mascaro et al., 2021). Compared with hard voting, these methods preserve more information during fusion and generally produce a more informative 3D estimate.

Visibility-aware filtering, view selection, and learned aggregation. The literature also contains methods that account for view quality more explicitly, although often through filtering or selection rather than learned continuous weighting. Dense image predictions may be back-projected onto a 3D mesh, with visibility constraints and geometric consistency enforced during semantic fusion (Rong et al., 2022). Other approaches cast the problem as an *optimal view selection* task, in which the most suitable image for 2D feature extraction is chosen based on visibility and projected support criteria (Adam et al., 2019). Closely related work has also emphasized the difficulty of transferring image evidence to 3D under heterogeneous viewing conditions in multi-view point labeling pipelines (Peters et al., 2023). Beyond post-hoc semantic transfer, joint image-3D segmentation architectures have further explored learned view aggregation. In this setting, attention-based multi-view aggregation can be used to merge image features according to viewing conditions for large-scale 3D semantic segmentation (Robert et al., 2022). Other multimodal architectures instead aggregate image features into 3D point representations earlier in the pipeline, for example through multi-view feature lifting and point-based fusion (Jaritz et al., 2019), projection-based multimodal fusion for LiDAR segmentation (Alnaggar et al., 2021), or recent multi-view-guided LiDAR segmentation networks that integrate image information directly into the backbone (Liu et al., 2024).

Positioning of DeepChoice. Although our method is closest in spirit to the last group, it targets a different regime. Unlike end-to-end image-3D fusion architectures, DeepChoice operates in a *post-hoc* image-guided pipeline: the upstream 2D segmenter is kept fixed, and the model learns only how much each valid view should contribute to the final 3D semantic decision. In contrast to hard voting, weighted label selection, or visibility-based filtering alone, we learn a *continuous normalized weight* for each valid point-view observation from compact geometric and image-quality descriptors, and use these weights to aggregate per-view semantic probability vectors directly at the 3D point level.

3. Method

3.1 Problem Formulation

We consider image-guided multi-view 3D semantic segmentation, where a point cloud is observed by multiple calibrated images. An upstream 2D segmentation network with a Swin-Tiny backbone (Liu et al., 2021) and a UPerNet decoder (Xiao et al., 2018) first predicts, for each image pixel, a class-probability vector over C semantic classes. These 2D predictions are then reprojected onto the 3D point cloud and fused across views.

For each 3D point p_i, the preprocessing stage identifies a set of valid image observations

$$\mathcal{V}_i = \{v_{i,1}, \ldots, v_{i,N_i}\}, \tag{1}$$

where N_i denotes the number of retained views for point p_i, with $1 \leq N_i \leq K$, and K a fixed maximum number of views. In the default setting, we retain at most $K = 10$ views per point.

Each valid observation $\mathbf{v}_{i,j}$ is associated with a per-view class-probability vector $\mathbf{s}_{i,j} \in [0,1]^C$, with $\sum_{c=1}^{C} s_{i,j}^{(c)} = 1$, produced by the upstream 2D segmenter, and a visibility descriptor $\mathbf{c}_{i,j} \in \mathbb{R}^6$ characterizing the expected reliability of that observation.

In our default formulation, the visibility descriptor contains six criteria:

$$\mathbf{c}_{i,j} = [\text{angle}, \text{distance}, \text{contrast}, \text{blur}, \text{SNR}, \text{saturation}]. \tag{2}$$

The role of DeepChoice is not to re-predict semantics from raw image content, but to learn how to weight already available 2D semantic predictions according to the expected reliability of each observation. In some experimental variants, we additionally concatenate the per-view class-probability vector $\mathbf{s}_{i,j}$ to the visibility descriptor in order to test whether view weighting benefits from class-dependent semantic context. The output of the model, however, remains unchanged in all cases: DeepChoice always predicts one scalar weight per view. Figure 1 summarizes the overall pipeline.

3.2 Visibility-Aware View Representation

The preprocessing stage is performed offline and produces precomputed point-wise observations used for training and inference. For each tile, 3D points are first associated with estimated normals. For each camera, points are then projected into the image plane, and validated through a depth-consistency visibility test.

For each visible 3D point–pixel pair, the pipeline extracts:

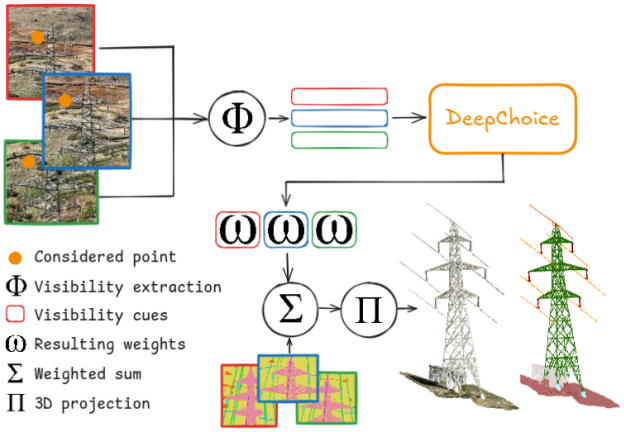


Figure 1. DeepChoice pipeline: for each point, extract visibility cues per view, predict per-view weights, fuse per-view class probabilities to point-wise probabilities.

- class-probability vector,
- geometric visibility criteria (incidence angle and point-to-camera distance),
- image-quality criteria (local contrast, blur, SNR, saturation).

Before training, visibility criteria are normalized using predefined minimum and maximum values. For each point, at most K views are retained using a deterministic ranking strategy based on camera-to-point distance. The resulting samples contain tensors of shape $[K, 6]$ for visibility descriptors and $[K, C]$ for per-view class probabilities, along with a Boolean mask that distinguishes valid observations from zero-padded entries introduced to maintain a fixed size along the K-view dimension.

When probability scores are included as additional inputs, they are concatenated to the visibility descriptor. This produces an augmented per-view descriptor of dimension $F+C$, while leaving the fusion objective itself unchanged. The motivation is that the most informative viewing conditions may differ across semantic categories: for instance, the cues that best characterize a reliable observation of an insulator may differ from those that are most useful for a pylon or for vegetation. By providing both visibility cues and class-probability context, the weighting network can learn class-dependent view-selection behavior while still predicting a single scalar importance weight per observation.

3.3 Late Fusion by Learned View Weighting

Given the retained observations of point p_i, DeepChoice predicts one scalar score per view from the corresponding per-view descriptors:

$$\mathbf{w}_i = f_\theta(\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,K}), \tag{3}$$

where f_θ is a lightweight neural network, $\mathbf{x}_{i,j} = \mathbf{c}_{i,j}$ in the default formulation, and $\mathbf{x}_{i,j} = [\mathbf{c}_{i,j}, \mathbf{s}_{i,j}]$ in the score-augmented variants. Invalid padded views are masked out.

The normalized fusion weights are obtained through a masked softmax over valid views:

$$\alpha_{i,j} = \frac{\exp(w_{i,j})}{\sum_{k \in \mathcal{M}_i} \exp(w_{i,k})}, \qquad j \in \mathcal{M}_i, \tag{4}$$

The fused per-point class distribution is then computed as

$$\hat{\mathbf{p}}_i = \sum_{j \in \mathcal{M}_i} \alpha_{i,j} \mathbf{s}_{i,j}, \tag{5}$$

and the final class prediction is

$$\hat{y}_i = \arg\max_c \hat{\mathbf{p}}_i^{(c)}. \tag{6}$$

This formulation makes the model explicitly responsible for estimating view usefulness rather than relearning semantic evidence from raw imagery. Since the upstream 2D predictions are already normalized class distributions, fusion is performed directly in probability space.

3.4 Model Variants

We study two variants of the weighting function f_θ.

MLP. The first variant concatenates the K per-view descriptors into a single vector and predicts K view scores with a multilayer perceptron. Depending on the experimental setting, each descriptor contains either visibility cues only or the concatenation of visibility cues and per-view softmax scores. This variant ignores explicit cross-view interactions and serves as a lightweight baseline.

Transformer. The second variant treats the K observations as a sequence of tokens, each token corresponding to one per-view descriptor. A compact Transformer encoder first projects each token to a latent embedding, then applies self-attention across views, and finally outputs one scalar score per token. In the reported configuration, the encoder uses two Transformer layers, one attention head, a model dimension of 32, and a feed-forward dimension of 64. This design enables the model to consider all available observations of a point together, adapting the importance of each observation according to the presence of the others.

In the score-augmented setting, each token contains $F+C$ input values. In our GridNet-HD experiments, this corresponds to $6 + 11 = 17$ dimensions per view. The Transformer therefore processes 17-dimensional tokens, whereas the MLP receives a flattened input of dimension $K \times 17$, i.e. $10 \times 17 = 170$ in the default setting with $K = 10$.

3.5 Training Objective

Let y_i be the ground-truth class of point p_i. Training is performed on the fused per-point class distribution:

$$\mathcal{L} = -\sum_i \log \hat{\mathbf{p}}_i^{(y_i)}. \tag{7}$$

In practice, we use a negative log-likelihood (NLL) loss on the fused probability vectors rather than a standard cross-entropy loss on logits. This follows directly from the structure of the pipeline: the upstream 2D network outputs per-view softmax probabilities, and DeepChoice computes a combination of these probabilities. The fused output is therefore already a valid class distribution. Supervision is applied only at the fused point level; the model is never given explicit target weights for individual views, but instead learns a view-weighting policy solely through the final point-classification objective.

4. Experiments

4.1 Scope of the Experimental Study

Our experiments are designed to answer four questions: (i) whether learned visibility-aware weighting improves over non-learned fusion baselines, (ii) whether a Transformer-based weighting model is preferable to a simpler MLP, (iii) whether augmenting visibility cues with per-view semantic score vectors improves view weighting, and (iv) how performance varies with the number of retained views.

4.2 Experimental Setup

Datasets. We evaluate DeepChoice on two datasets. *GridNet-HD* is our main benchmark and is used for final test-set comparison as well as validation-set ablations. We follow the split protocol described in the *GridNet-HD* paper (Carreaud et al., 2026a). The dataset provides co-registered LiDAR point clouds (2.5 billion), calibrated RGB images (7,700), and camera poses. The dataset and benchmark resources are publicly available online[1].

As a complementary evaluation set, we use the *Images&PointClouds Cultural Heritage* dataset (Pellis et al., 2025b). This second benchmark differs significantly from *GridNet-HD* in terms of its acquisition setup, scene structure and semantic content. It is therefore used to evaluate the performance of the proposed fusion strategy in a substantially different domain. Due to the small number of scenes and their strong visual heterogeneity, a strict cross-scene image-segmentation protocol produces extremely weak 2D logits on some splits and prevents a meaningful analysis of the fusion stage. We therefore adopt the following protocol. The upstream 2D image segmenter is trained on scenes `1_SC`, `2_OSA`, and `3_SS` in order to obtain usable semantic score maps, while DeepChoice itself is trained only on scenes `1_SC` and `2_OSA` and evaluated on scene `3_SS`. In other words, the 2D segmenter has been made sufficiently stable to provide informative logits, while the fusion model is still being evaluated using a scene that has been removed from the training set. This setting should therefore be interpreted as a cross-domain evaluation of the DeepChoice fusion mechanism on a different dataset, rather than as a strict cross-dataset transfer. Accordingly, this experiment does not evaluate zero-shot generalization of the full image-guided pipeline, since the upstream 2D segmenter has been exposed to the target scene during training. Figure 2 illustrates the visual gap between the two datasets.

Each scene in *Images&PointClouds Cultural Heritage* provides RGB images, camera intrinsics and extrinsics, a transformed point cloud, and 2D labels derived from 3D reprojection. We keep the released annotations and camera files for comparability, while explicitly accounting for the fact that the projected 2D labels contain visible local artifacts and boundary inaccuracies in some views.

In addition, the raw point-cloud files of *Images&PointClouds Cultural Heritage* are not fully homogeneous across scenes. Some scenes provide precomputed normals directly in the point-cloud text files, while others omit normals. In our preprocessing pipeline, we therefore parse the scene-specific format explicitly and recompute normals only when normals are not available in the released files.

[1] `https://huggingface.co/collections/heig-vd-geo/gridnet-hd`

(a) *Images&PointClouds Cultural Heritage*: terrestrial heritage scene.

(b) *GridNet-HD*: outdoor power-line scene.

Figure 2. Visual comparison of the two evaluation datasets: (a) a terrestrial heritage scene from *Images&PointClouds Cultural Heritage*; (b) an outdoor electrical-infrastructure scene from *GridNet-HD*.

Preprocessing and training. For each 3D point, we compute normals, project the point into the available images, estimate visibility through depth consistency, and retain up to $K = 10$ views using the same distance-based ranking strategy described in Section 3. All models are trained on precomputed batch files containing visibility descriptors, per-view class-probability vectors, masks, and point labels. The training objective is the NLL loss on fused point-wise class distributions. Optimization is performed with AdamW. Unless otherwise stated, all quantitative results are reported using mean Intersection over Union (mIoU), mean F1-score (mF1), and per-class IoU.

Evaluation baselines. We compare DeepChoice with three non-learned fusion baselines:

- **Mean-probability baseline.** For each point, the per-view class-probability vectors are averaged over valid views and the final class is obtained with an `argmax`.
- **Hard-vote baseline.** Each view contributes only its winning class, and the final prediction is obtained by majority vote in label space.
- **AnyView oracle.** A point is counted as correctly classified if at least one valid view predicts the correct class. This baseline is not a deployable method, but an upper-bound style reference indicating how often the correct semantic label is already available in at least one view.

Method	mIoU (%)	Params
Image-guided		
Hard Vote	66.78	60 M
Mean Prob.	69.37	60 M
AnyView oracle	84.33	60 M
DeepChoice-MLP	70.52	60 M + 0.09 M
DeepChoice-Transformer	**70.63**	60 M + 0.02 M
3D-only		
SPT (XYZ+RGB)	66.90	0.21 M
PTv3 (XYZ+RGB, 10 cm)	64.53	46.2 M
PTv3 (XYZ+RGB, 5 cm)	66.86	46.2 M
+ TTA	69.32	46.2 M

Table 1. mIoU comparison on the *GridNet-HD* test split. The first block reports image-guided methods with total parameter counts; the second reports 3D-only baselines from the *GridNet-HD* paper.

4.3 GridNet-HD Global Results

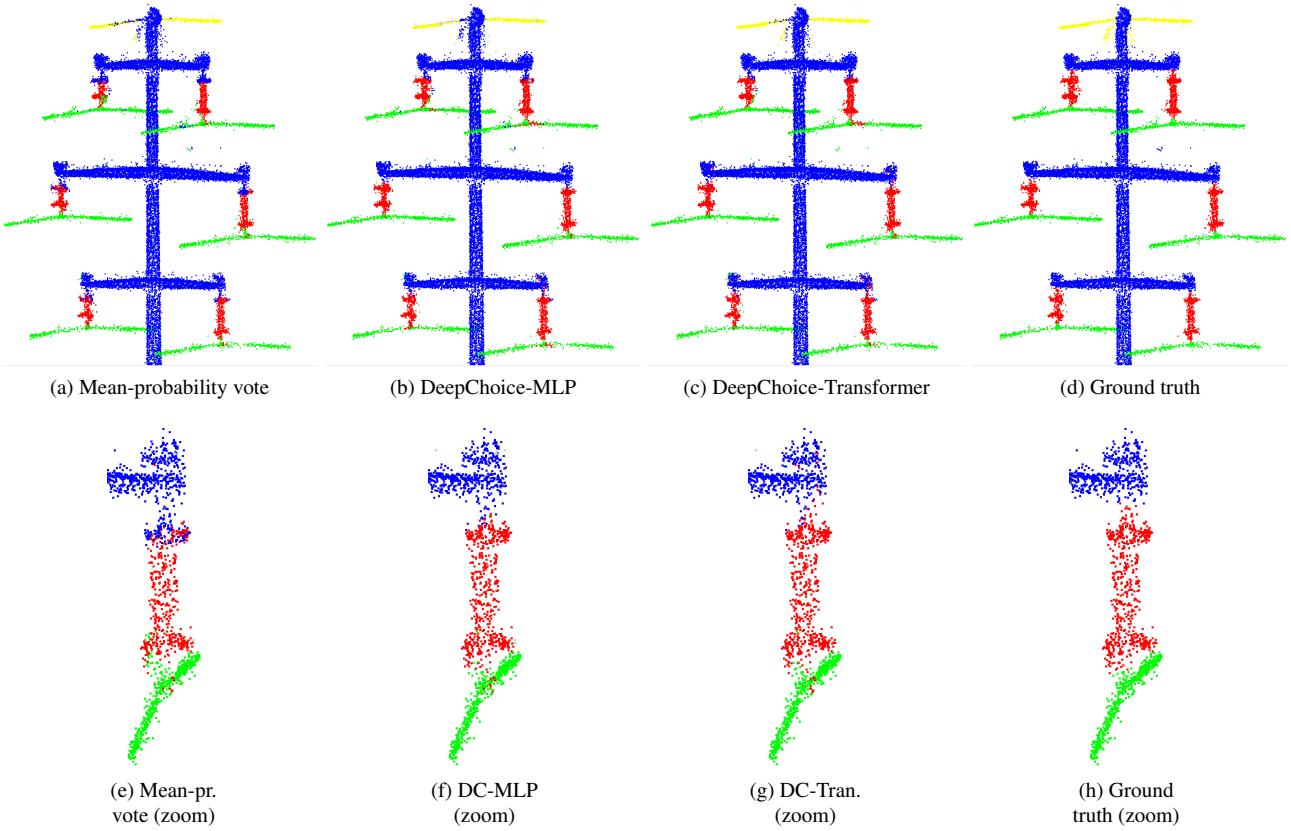
We first compare the proposed learned fusion models against the non-learned baselines on the *GridNet-HD* test split. Hyperparameters and feature configurations are selected on the validation set, and only the selected configurations are reported on test. Table 1 summarizes the mIoU comparison across image-guided fusion baselines, the two DeepChoice variants, and representative 3D-only methods reported in the *GridNet-HD* paper (Carreaud et al., 2026a). On a single A40 GPU, descriptor extraction and projection require about 21 s per million points, and DeepChoice inference about 6 s per million points. This additional cost remains moderate, as the module operates on compact per-view descriptors and adds at most 0.09 M parameters.

Both DeepChoice variants outperform the non-learned image-guided baselines. DeepChoice-Transformer improves mIoU by 3.85 points over hard voting and by 1.26 points over mean-probability fusion, while DeepChoice-MLP yields a very similar gain. This indicates that the main improvement comes from replacing fixed fusion rules with learned view weighting. Although the Transformer only slightly improves over the MLP in absolute mIoU, it does so with far fewer parameters, making it the most parameter-efficient fusion variant overall.

The gap to the *AnyView* oracle remains substantial, indicating that view selection and weighting are still not fully solved. In particular, the oracle suggests that the correct class is already present in at least one valid view much more often than the current learned models are able to recover, meaning that the current handcrafted visibility descriptors are informative but not sufficient to fully exploit the best available view evidence.

The comparison with 3D-only baselines should be interpreted with caution, since the image-guided and 3D-only methods rely on different input modalities and supervision pipelines. Their inclusion is intended to place the reported results within the performance range of current alternatives on the same benchmark rather than to claim strict architectural comparability. Within this context, DeepChoice-Transformer achieves the highest mIoU reported in Table 1.

Figure 3 complements these results with a qualitative example on *GridNet-HD*. Compared with mean-probability fusion, both learned variants better preserve thin structures and reduce local label confusion, with the Transformer producing the closest match to the ground truth in the zoomed region.



(a) Mean-probability vote (b) DeepChoice-MLP (c) DeepChoice-Transformer (d) Ground truth

(e) Mean-pr. vote (zoom) (f) DC-MLP (zoom) (g) DC-Tran. (zoom) (h) Ground truth (zoom)

Figure 3. Qualitative comparison on a *GridNet-HD* scene. The first row shows, in order, mean-probability fusion, DeepChoice-MLP, DeepChoice-Transformer, and the ground truth. The second row reports the corresponding zoomed-in views in the same order.

4.4 GridNet-HD Per-Class Results

Table 2 reports the per-class IoU on *GridNet-HD* test split. This analysis is important because the benefit of visibility-aware fusion is expected to vary across categories depending on their geometry, image appearance, and sensitivity to view obliqueness.

The per-class results show that the gain is not uniform across categories. The strongest improvements are observed for *Conductor*, *Pylon*, and *Structure*, which are precisely the classes for which view quality is most critical. These categories correspond to thin, elongated, or structurally complex objects that are especially sensitive to oblique views, long distances, and local image degradation. On such classes, replacing mean-probability fusion with learned weighting yields large gains.

More modest improvements are observed for vegetation and road-related classes. This is consistent with the fact that these classes are often represented by larger and more redundant image evidence, for which fixed averaging is already effective. The most notable failure case is *Water*, for which both learned variants underperform mean-probability fusion, with the Transformer dropping particularly strongly. We attribute this result primarily to the extreme under-representation of this class in the dataset, which limits the ability of the weighting network to learn robust fusion behavior for such observations. This highlights a limitation of the current approach: while DeepChoice is especially beneficial for difficult thin structures, its gains are not uniform across all semantic categories. This also suggests that learned weighting may be more sensitive than simple averaging on rare classes with limited supervision.

4.5 Cultural Heritage Results

We also report results on the *Images&PointClouds Cultural Heritage* dataset, which differs markedly from *GridNet-HD* in acquisition setup and scene structure while preserving the same image-guided 2D-to-3D fusion principle. For this dataset, we retain only the best-performing DeepChoice input configuration identified on *GridNet-HD*, namely the combination of per-view semantic scores and all visibility criteria, and compare its MLP and Transformer variants against hard-vote fusion, mean-probability fusion, and the AnyView oracle (Table 3).

The gains on *Images&PointClouds Cultural Heritage* are smaller than on *GridNet-HD*, but they remain informative. In this setting, DeepChoice-Transformer improves over hard voting by 1.55 mIoU points and slightly surpasses mean-probability fusion by 0.48 mIoU points. This suggests that the proposed weighting strategy retains some benefit under a different acquisition regime and on a very different dataset.

The smaller margin also indicates that this setting is more challenging than *GridNet-HD*. This likely stems from the very small number of available scenes, their strong inter-scene heterogeneity, and the fact that the released image labels are affected by reprojection artifacts and boundary inaccuracies. Overall, these results suggest that DeepChoice remains applicable beyond the main benchmark, while also highlighting that its benefit depends on the quality, consistency, and diversity of the underlying data.

Class	Hard Vote (%)	Mean Prob. (%)	DeepChoice-MLP (%)	DeepChoice-Trans. (%)	AnyView (%)
Pylon	82.87	86.77	92.12	92.85	95.03
Conductor	55.68	66.03	75.82	77.79	88.08
Structure	40.01	47.72	51.93	54.93	78.33
Insulator	73.74	75.84	76.95	76.98	93.72
High veg.	83.15	83.74	84.08	84.16	90.58
Low veg.	60.46	61.19	60.18	60.32	74.71
Herb. veg.	83.47	83.85	84.37	84.43	88.47
Soil	38.37	39.61	42.82	43.25	53.60
Road	74.47	74.68	75.14	76.35	88.88
Water	75.52	77.24	66.43	59.13	90.66
Building	66.85	66.40	65.91	66.68	85.61
mIoU	**66.78**	**69.37**	**70.52** $^{\uparrow 1.15}$	**70.63** $^{\uparrow 1.26}$	**84.33**

Table 2. Per-class IoU comparison on the *GridNet-HD* test split.

Method	mIoU (%)	mF1 (%)
Hard Vote	65.66	73.61
Mean Prob.	66.73	74.60
AnyView	77.95	83.06
DeepChoice-MLP	66.35	74.54
DeepChoice-Transformer	67.21	75.49

Table 3. Results on the 3_SS scene of *Images&PointClouds Cultural Heritage* dataset. The two DeepChoice rows report the configuration that combines per-view semantic scores with all visibility criteria.

4.6 Ablation Study

All ablations are conducted on the *GridNet-HD* validation split. We first analyze which inputs should be provided to the weighting network, and then study the effect of the maximum number of retained views.

4.6.1 Input Ablation

Protocol. Our first question is whether adding the per-view softmax score vector to the weighting-network input improves performance beyond visibility cues alone. The underlying hypothesis is that optimal viewing conditions may not be identical across semantic classes: a view that is reliable for vegetation may not be equally informative for thin structures such as cables or insulators. We therefore evaluate five feature configurations. The first uses visibility criteria only, which corresponds to the default formulation of DeepChoice. The second augments these criteria with the per-view semantic score vector. We then evaluate two partial hybrids, one combining semantic scores with image-quality criteria only, and one combining semantic scores with geometric criteria only. Finally, we consider a semantic-scores-only variant, which removes explicit visibility cues altogether and tests whether semantic confidence alone is sufficient for view weighting.

Quantitative results. Table 4 shows that the best configuration combines semantic scores with all visibility criteria for both architectures. This result supports the hypothesis that semantic probabilities alone are not sufficient to characterize view reliability: the weighting network also benefits from explicit geometric and image-quality cues. Conversely, using visibility cues alone or semantic scores alone leads to lower performance, which indicates that the two sources of information are complementary. The partial variants further clarify this result. Combining semantic scores with only image criteria or only geometric criteria improves over the visibility-only baseline, but both remain below the full combination. This suggests that neither geometric visibility nor image quality alone is sufficient to explain view usefulness. Instead, the best weighting policy requires both an explicit description of observation conditions and semantic context from the image segmenter. This is consistent with the intended role of DeepChoice: the same geometric configuration may not have the same relevance for all classes, and the semantic score vector helps the model adapt its weighting policy accordingly.

Qualitative interpretation of the learned weights. Figure 4 provides a qualitative view of the learned weighting output for the best-performing configuration. The trends in Figure 4 are consistent with the intended behavior of a visibility-aware fusion model. In particular, the learned weight decreases for large incidence angles and tends to decrease with view-to-point distance, which matches the intuition that oblique and distant observations are less reliable. The remaining criteria exhibit smoother and more context-dependent effects. Local contrast and blur show gradual trends, while SNR and saturation display more irregular behavior and broader interquartile ranges, suggesting that these descriptors are less directly interpretable in isolation. Taken together, these curves indicate that the weighting policy is not governed by a single dominant feature, but by the combination of several complementary indicators. This interpretation is further supported by the class-wise analyses, which reveal different effects across semantic categories and likely explain part of the observed variability. Overall, these results suggest that richer descriptors could further improve view selection.

4.6.2 Influence of the Number of Retained Views

Protocol. We finally study the effect of the maximum number of retained views. This second ablation is conducted on the *GridNet-HD* validation split using the best input configuration identified above, namely the combination of semantic scores and all visibility criteria. The corresponding comparison is reported in Table 5.

Results and interpretation. Increasing the number of retained views consistently improves performance for both architectures. For the Transformer, mIoU rises from 76.60 at $K = 3$ to 79.65 at $K = 10$; for the MLP, it rises from 77.40 to 79.13 over the same range. This suggests that the gain does not come only from selecting a few strong observations, but also from exploiting a richer multi-view context once the weighting model is able to suppress weak views.

Input variant	Transformer mIoU (%)	Transformer mF1 (%)	MLP mIoU (%)	MLP mF1 (%)
Visibility criteria only	78.60	87.03	77.17	86.01
Semantic scores + all visibility criteria	79.65	87.72	79.13	87.33
Semantic scores + image criteria only	79.03	87.25	78.97	87.19
Semantic scores + geometric criteria only	79.15	87.37	78.93	87.17
Semantic scores only	78.71	87.02	78.83	87.10

Table 4. Validation-set ablation study for the weighting-network inputs.

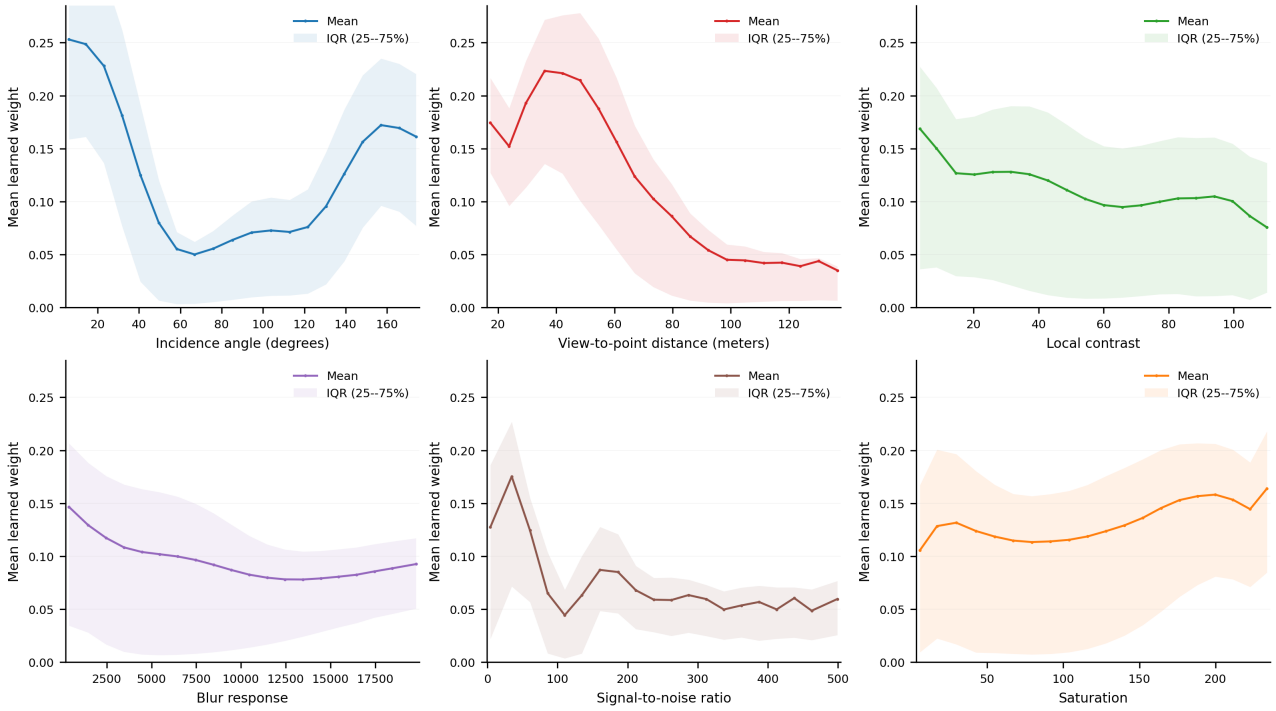


Figure 4. Analysis of the learned weighting output for the best-performing DeepChoice configuration. Each subplot shows the distribution of learned view weights when observations are grouped according to one visibility criterion.

Method and K	mIoU (%)	mF1 (%)
DeepChoice-Transformer, $K = 3$	76.60	85.56
DeepChoice-Transformer, $K = 5$	78.56	86.95
DeepChoice-Transformer, $K = 10$	79.65	87.72
DeepChoice-MLP, $K = 3$	77.40	86.15
DeepChoice-MLP, $K = 5$	78.91	87.19
DeepChoice-MLP, $K = 10$	79.13	87.33

Table 5. Validation-set ablation on the maximum number of retained views K, using the best input configuration (semantic scores + all visibility criteria).

5. Conclusion and Perspectives

This paper introduced **DeepChoice**, a lightweight view-weighting module for image-guided 3D semantic segmentation. DeepChoice learns to assign a relevance weight to each valid image observation of a 3D point and fuses per-view 2D semantic probabilities into a final 3D point-wise prediction. The method is sensor-agnostic, does not require meshing, and can be integrated into existing image-to-point transfer pipelines.

Experiments on the full *GridNet-HD* benchmark showed that learned view weighting improves over hard voting and mean-probability fusion. The best results were obtained by combining visibility cues with per-view semantic scores, which suggests that view reliability depends both on observation conditions and on semantic context. The compact Transformer variant provided the best accuracy-to-parameter ratio among the tested variants, while the largest gains were observed on thin and difficult classes such as conductors, pylons, and structures.

We further evaluated DeepChoice on the *Images&PointClouds Cultural Heritage* dataset, where the gains were smaller but still showed that the proposed weighting strategy remains beneficial in a substantially different acquisition setting.

Perspectives. Future work should evaluate DeepChoice on larger and more homogeneous terrestrial datasets with clearer train/validation/test splits, in order to better assess learned view weighting beyond the current benchmarks. Another important direction is to reduce the remaining gap to the *AnyView* oracle through richer observation descriptors, stronger weighting strategies, and a better understanding of class-dependent weighting behavior. Finally, it would be valuable to study the impact of stronger upstream 2D segmenters, since the fused 3D prediction directly depends on the quality of the per-view semantic logits. This includes both context-aware architectures tailored to large images, such as CASWiT (Carreaud et al., 2026b), and lighter alternatives such as SegFormer (Xie et al., 2021), which may offer a better trade-off between semantic quality and computational efficiency.

References

Adam, A., Grammatikopoulos, L., Karras, G., Protopapadakis, E., Karantzalos, K., 2019. A semantic 3d point cloud segmentation approach based on optimal view selection for 2d image feature extraction. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W17, 9–14.

Alnaggar, Y. A., Afifi, M., Amer, K., ElHelw, M., 2021. Multi projection fusion for real-time semantic segmentation of 3d lidar point clouds. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1800–1809.

Carreaud, A., Li, S., De Lacour, M., Skaloud, J., Gressin, A., 2025. Advanced image-based methods for 3d semantic segmentation of lidar point clouds in electrical infrastructure applications. *IGARSS - IEEE International Geoscience and Remote Sensing Symposium*.

Carreaud, A., Li, S., Lacour, M. D., Frinde, D., Skaloud, J., Gressin, A., 2026a. Gridnet-hd: A high-resolution multi-modal dataset for lidar-image fusion on power line infrastructure.

Carreaud, A., Mariani, F., Gressin, A., 2022. Automating the underground cadastral survey: A processing chain proposal. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2022, 565–570.

Carreaud, A., Naha, E., Chansel, A., Lahellec, N., Skaloud, J., Gressin, A., 2026b. Context-aware semantic segmentation via stage-wise attention.

Genova, K., Yin, X., Kundu, A., Pantofaru, C., Cole, F., Sud, A., Brewington, B., Shucker, B., Funkhouser, T., 2021. Learning 3d semantic segmentation with only 2d image supervision. *International Conference on 3D Vision (3DV)*.

Hermans, A., Floros, G., Leibe, B., 2014. Dense 3d semantic mapping of indoor scenes from rgb-d images. *IEEE International Conference on Robotics and Automation (ICRA)*.

Huang, B., Wang, Z., Chen, J., Zhou, B., Ma, H., 2025. A segmentation method for LiDAR point clouds of aerial slender targets. *Frontiers in Physics*, Volume 13 - 2025.

Jaritz, M., Gu, J., Su, H., 2019. Multi-view pointnet for 3d scene understanding. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

Jeon, J., Jung, J., Kim, J., Lee, S., 2018. Semantic Reconstruction: Reconstruction of Semantically Segmented 3D Meshes via Volumetric Semantic Fusion. *Computer Graphics Forum*, 37(7), 25–35.

Kundu, A., Yin, X., Fathi, A., Ross, D., Brewington, B., Funkhouser, T., Pantofaru, C., 2020. Virtual multi-view fusion for 3d semantic segmentation. A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (eds), *Computer Vision – ECCV 2020*, Springer International Publishing, Cham, 518–535.

Liu, Y., Liu, Y., Duan, Y., 2024. MVG-Net: LiDAR Point Cloud Semantic Segmentation Network Integrating Multi-View Images. *Remote Sensing*, 16(15).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Mascaro, R., Teixeira, L., Chli, M., 2021. Diffuser: Multi-view 2d-to-3d label diffusion for semantic scene segmentation. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 13589–13595.

McCormac, J., Handa, A., Davison, A. J., Leutenegger, S., 2017. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. *IEEE International Conference on Robotics and Automation (ICRA)*, 4628–4635.

Pellis, E., Masiero, A., Betti, M., Tucci, G., Grussenmeyer, P., 2025a. A Deep Learning Multiview Approach for the Semantic Segmentation of Heritage Building Point Clouds. *International Journal of Architectural Heritage*, 0(0), 1–23.

Pellis, E., Masiero, A., Betti, M., Tucci, G., Grussenmeyer, P., 2025b. A photogrammetric image-point dataset for the semantic segmentation of heritage buildings. *Data in Brief*, 60, 111661.

Pellis, E., Murtiyoso, A., Masiero, A., Tucci, G., Betti, M., Grussenmeyer, P., 2022. 2d to 3d label propagation for the semantic segmentation of heritage building point clouds. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2022, 861–868.

Peng, S., Genova, K., Jiang, C. M., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., 2023. Openscene: 3d scene understanding with open vocabularies. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–824.

Peters, T., Brenner, C., Schindler, K., 2023. Semantic Segmentation of Mobile Mapping Point Clouds via Multi-View Label Transfer. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202, 30–39.

Robert, D., Vallet, B., Landrieu, L., 2022. Learning multi-view aggregation in the wild for large-scale 3d semantic segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5565–5574.

Rong, M., Cui, H., Hu, Z., Jiang, H., Liu, H., Shen, S., 2022. Active Learning Based 3D Semantic Labeling From Images and Videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12), 8101–8115.

Stathopoulou, E.-K., Remondino, F., 2019. Semantic Photogrammetry – Boosting Image-Based 3D Reconstruction with Semantic Labeling. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W9, 685–690.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J., 2018. Unified perceptual parsing for scene understanding. *Proceedings of the European Conference on Computer Vision (ECCV)*.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P., 2021. Segformer: Simple and efficient design for semantic segmentation with transformers. M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (eds), *Advances in Neural Information Processing Systems*, 34, Curran Associates, Inc., 12077–12090.