EN RU PROGRAM SYSTEMS: THEORY AND APPLICATIONS ISSN 2079-3316
Research Article ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

UDC 004.932.75'1, 004.89
10.25209/2079-3316-2026-17-2-191-262

Roof-DeGAN: a hybrid GAN with cross-scale attention for aerial roof inpainting

Igor Victorovich **Vinokurov**[1]✉, Georgy Mikhailovich **Lapankov**[2], Georgy Dmitrievich **Umarov**[3]

[1–3] Financial University under the Government of the Russian Federation, Moscow, Russia

[1]✉ *igvvinokurov@fa.ru*

Abstract. This paper proposes a hybrid generative adversarial network, Roof-DeGAN, for restoring damaged and missing roof areas in aerial images. The architecture combines a densely connected Vision Transformer in the generator with a multi-scale discriminator featuring cross-scale attention. The model integrates the advantages of GANs, diffusion modeling elements, and transformer mechanisms. Experiments on real data from the PLC «Roscadastr» demonstrate that Roof-DeGAN outperforms existing methods, achieving PSNR = 33.7 dB, SSIM = 0.971, LPIPS = 0.048, and FID = 17.8 with an inference time of 0.15 seconds per 256×256 image. The developed approach shows high practical value for cadastre maintenance and cartographic data updating tasks. (*Linked article texts in English and in Russian*).

Key words and phrases: generative adversarial networks, Roof-DeGAN, image inpainting, aerial imagery, remote sensing, roof reconstruction

2020 *Mathematics Subject Classification*: 68T20; 68T07, 68T45

For citation: Igor V. Vinokurov, Georgy M. Lapankov, Georgy D. Umarov. *Roof-DeGAN: a hybrid GAN with cross-scale attention for aerial roof inpainting*. Program Systems: Theory and Applications, 2026, **17**:2(71), pp. 191–262. (*In English, in Russian*). https://psta.psiras.ru/read/psta2026_2_191-262.pdf

Introduction

The restoration of images and building contours from aerial photographs for creating up-to-date terrain maps and cadastral assessment of capital construction objects is one of the main tasks addressed by the PLC «Roscadastr». Effective solution of this problem is of significant interest for cartography, urban infrastructure monitoring, and urban planning [1, 2]. However, in practice, aerial photographs often contain defects caused by atmospheric factors (cloudiness, fog), temporary objects (construction equipment, vehicles, tree canopies), or technical limitations of imaging (sensor noise, low resolution). The quality of restoration directly affects the accuracy of subsequent automated analysis, including semantic and instance segmentation, 3D model construction, and condition assessment of capital construction objects.

Traditional image restoration methods, such as Navier-Stokes inpainting [3] and patch-matching based on similarity search [4], demonstrate limited effectiveness when dealing with complex roof structures. These approaches rely predominantly on low-level features (pixel intensity, color and structure gradients) and do not account for object semantics, leading to blurred boundaries, distortion of geometric shapes (roof slopes, ridges, valleys), and the appearance of unnatural texture artifacts. When reconstructing large damaged areas, such methods cannot generate fundamentally new content, which is critical for aerial photographs with large defects.

Modern deep learning methods offer three main directions for solving this problem: generative adversarial networks (GANs), diffusion probabilistic models, and their hybrid combinations.

Generative adversarial networks remain in demand due to their high inference speed and robustness to limited training data. The foundations of adversarial training were presented in [5], and conditional GANs (cGANs) for image-to-image translation have become a basic approach to restoration. However, classical architectures such as Pix2Pix [6] and ESRGAN [7] show degraded quality when processing large masks due to the limited capacity of the discriminator and insufficient penalty for high-frequency distortions, which is particularly noticeable on roofing material textures.

Diffusion probabilistic models demonstrate high generation quality. The base DDPM architecture [8] and its development for restoration tasks [9] achieve high realism but require thousands of inference iterations and large amounts of data. In 2025, specialized adaptations for remote sensing appeared: SatDiff [10] based on Stable Diffusion provides high-quality satellite image restoration; the KAO method [11] introduces kernel-adaptive optimization, outperforming previous approaches on structural tasks; the Image Characteristic-Guided method [12] accounts for low-rank image properties. Despite superior metrics, these models retain high computational complexity and low inference speed.

Hybrid methods combine the advantages of GANs and diffusion. Modern hybrid GAN implementations, such as DeGAN [13], BD-VITGAN [14], and TAMGAN [15], integrate transformers, dense connections, and multi-level attention, significantly improving structural consistency and perceptual quality in remote sensing tasks. Other approaches, such as DSEPGAN [16], achieve a balance between speed, determinism, and restoration quality, which is particularly important for specialized aerial roof imaging tasks.

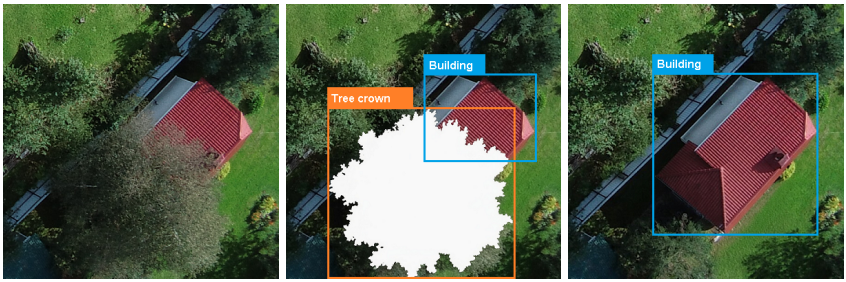
The current state of the field is characterized by fragmentation: most methods are developed for general datasets; diffusion models require excessive resources; and hybrid approaches are insufficiently adapted to the specific characteristics of aerial roof images. There is no specialized method that combines robustness to small sample sizes, deterministic inference, high perceptual quality, and moderate computational complexity. The present work aims to address these shortcomings and continues the research described by the author in [17, 18]. A hybrid approach is proposed for restoring damaged roof areas in aerial photographs, based on a conditional generative adversarial network DeGAN with transformer blocks in the generator and a multi-level discriminator with cross-level attention. The composite loss function combines adversarial components, pixel-wise comparison, perceptual features, and diffusion enhancement, achieving a balance between geometric reconstruction accuracy and naturalness of roofing material textures.

Section 1 presents a formal problem statement for the restoration of missing roof areas. Section 2 describes the architecture of Roof-DeGAN, including the structure of the generator with transformer blocks (equipped with a dynamic attention sparsification mechanism) and a multi-scale discriminator, as well as the composite loss function. Section 3 presents the datasets used (ZRG for pre-training and the target dataset with synthetic damage), data preprocessing methods, as well as details of the two-stage model training (pre-training for segmentation and fine-tuning for restoration) and a list of quantitative evaluation metrics (PSNR, SSIM, LPIPS, FID). Section 4 analyzes the training dynamics and model convergence. Section 5 provides a quantitative comparison of the proposed approach with state-of-the-art methods. Section 6 presents an ablation study to evaluate the contribution of each component of the loss function, as well as the impact of pre-training on ZRG. Section 7 analyzes the restoration quality as a function of the damage area size. Section 8 discusses the limitations and promising directions for future research. The paper concludes with Section 9 and a list of references.

1. Problem Statement

The problem of restoring damaged or occluded areas of building roofs in aerial photographs can be formulated as a conditional image generation task. Let $I_{gt} \in \mathbb{R}^{H\times W\times 3}$ be the original undamaged RGB image of a building roof of size $H \times W$ pixels. A binary mask of the damaged region $M \in \{0,1\}^{H\times W}$ defines the pixels to be restored: $M_{ij} = 1$ corresponds to a damaged pixel, $M_{ij} = 0$ corresponds to intact regions. The damaged image is formed as $I_{in} = I_{gt} \odot (1 - M)$, where $\odot$ denotes element-wise multiplication.

The objective is to adjust the parameters θ of a parametric function $G_\theta(\cdot)$ such that, for the pair (I_{in}, M), it restores the complete image $I_{\text{rec}} = G_\theta(I_{\text{in}}, M)$, minimizing the generator loss function $\mathcal{L}_G$. A key feature of the problem is the need to generate new content while strictly preserving the geometric integrity of roof architectural elements (slopes, ridges, valleys) and the semantic consistency of roofing material textures (tiles, metal tiles, slate) with the surrounding context. This distinguishes it from general image inpainting tasks and requires specialized hybrid approaches (e.g., GANs with diffusion that account for geometric and semantic features).



(*a*) damaged image I_{in} (*b*) damage mask (I_{in}, M) (*c*) restored image I_{rec}

FIGURE 1. Example of restoring a damaged roof area

Figure 1 presents a sequence of three images illustrating the restoration process of an occluded roof area — the original damaged image, the damage mask (tree canopy), and the restoration result.

The necessity of solving this problem is driven by the practical needs of the PLC «Roscadastr». In the mass processing of aerial photographs for cadastral registration and updating of cartographic materials, a key step is the extraction of building contours and the formation of digital maps. However, the presence of damaged or occluded areas in the images (e.g., covered by tree canopies, cloud shadows, or temporary objects) causes standard segmentation and contour extraction algorithms to fail to correctly determine roof boundaries. An unreconstructed, noisy image does not allow obtaining a reliable binary object mask, which directly affects:

- the accuracy of contour maps — at the location of an occluded roof fragment, the contour is either broken or erroneously drawn along the boundary of the overlapping object (e.g., a tree canopy);
- the formation of digital elevation and terrain models — the height reference of building corners is distorted;
- cadastral registration — the area of the object is incorrectly calculated, leading to legal and economic consequences.

Manual processing of thousands of images with such defects is extremely labor-intensive and time-consuming. Developing an effective image restoration model allows automatic filling of occluded areas with semantically

consistent content while preserving roof geometry. As a result, mapping algorithms receive a complete image as input, ensuring correct contour extraction, area calculation, and the formation of reliable cartographic products. At the same time, processing time is reduced from several minutes per image to fractions of a second, which is critically important when working with regional and federal aerial photography archives.

2. Roof-DeGAN Architecture

The proposed model is a hybrid generative adversarial network in which the generator restores occluded image regions while the discriminator evaluates the plausibility of the resulting fragments. To improve restoration quality, modern components are employed: transformer blocks for capturing global context, dense connections for better feature propagation, and a diffusion enhancement component for increased stability and naturalness of textures.

The generator is built upon an architecture combining transformer blocks and dense convolutional connections, specifically adapted for satellite and aerial image restoration tasks. It consists of three main parts: a contracting path (encoder), a bottleneck layer, and an expanding path (decoder). Skip connectors between corresponding levels of the contracting and expanding paths transmit low-level features.

The contracting path consists of a sequence of blocks, each including a self-attention mechanism, a fully connected layer for capturing long-range dependencies in the image, and a dense convolutional block that performs the transformation:

$$y = \text{LeakyReLU}\left(\text{BN}\left(\text{Conv}_{3\times3}(x)\right)\right) + x,$$

where

$\text{Conv}_{3\times3}(\cdot)$ is a 2D convolution with a 3×3 kernel,
BN is batch normalization,
$\text{LeakyReLU}(\cdot)$ is an activation function with a negative slope.

The number of feature channels increases from 64 to 1024 (in the bottleneck), while the spatial resolution is halved at each level by using a convolution stride of 2.

The expanding path is symmetric to the contracting path and uses transposed convolutions to increase resolution. At each level, the following transformation is performed:

$$y = \text{TransposeConv}_{4\times4,\, s=2}\left(\text{Concat}\left(x_{dec}, x_{enc}\right)\right) + \text{ResidualBlock}(x),$$

where

$\text{TransposeConv}_{4\times4,s=2}(\cdot)$ is a transposed convolution with a 4×4 kernel and stride 2,

$\text{Concat}\left(x_{dec}, x_{enc}\right)$ is the concatenation (channel-wise joining) of features from the current level of the expanding path and the corresponding level of the contracting path via a skip connector,

$\text{ResidualBlock}(x)$ is an additional residual block for stabilizing training.

Skip connectors transmit detailed low-level features (boundaries, local textures), which prevents blurring of restored regions.

The generator's output layer consists of a 1×1 convolution with a $\tanh(\cdot)$ activation function, which maps pixel values to the range $[-1, 1]$:

$$\widehat{I} = \tanh\left(\text{Conv}_{1\times1}(z)\right),$$

where $\widehat{I}$ is the restored image and z is the output of the last block of the expanding path.

The discriminator is designed as a multi-level network with a cross-scale attention mechanism. It evaluates plausibility not on the entire image at once, but simultaneously at several levels of detail (from small fragments to large regions), linking information across different feature scales for more accurate assessment of restored fragment consistency. The discriminator's input is the concatenation of the damaged image and the damage mask. Its output is a set of probability maps at different resolutions. Each level includes a convolutional backbone with a $\text{LeakyReLU}(\cdot)$ activation function and spectral normalization, as well as a cross-scale attention module that links information between levels of detail. This approach provides accurate assessment of high-frequency details (roofing material textures) and roof geometric elements, while also improving training stability through gradient diversity.

Standard transformer blocks in the generator have quadratic computational complexity $O(n^2)$ and do not account for the specific characteristics of aerial roof images, which contain large homogeneous areas (flat slopes) alongside fine details (tiles, joints). To address this problem, Roof-DeGAN introduces a Dynamic Sparse Attention (DSA) mechanism. At each encoder level, for each patch, a local variability measure is computed:

$$v_i = \mathrm{Var}(p_i) + \mathrm{Var}(\nabla p_i),$$

where p_i are the pixel values in the i-th patch and ∇p_i is the gradient. For patches with low variability, attention is computed on a combination of:

- predictable neighboring patches (e.g., within a 3×3 window);
- 30% of randomly selected patches across the entire image.

Patches with high variability ($v_i \geq \tau$) are processed fully. The threshold τ is dynamically adjusted as a percentile of the distribution $\{v_i\}$ at the current level (with $\tau = P_{50}$, the 50th percentile, i.e., the median of the $\{v_i\}$ distribution). DSA reduces computational complexity from $O(n^2)$ to $O(n \cdot k)$, where $k \ll n$ is the effective attention size, and enables processing of higher-resolution images without quality loss on textured regions.

The organization of Roof-DeGAN is shown in Figure 2.

The overall loss function of the model $\mathcal{L}_G$ consists of several components:

$$\mathcal{L}_G = \lambda_{pix}\mathcal{L}_{pix} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{diff}\mathcal{L}_{diff} + \lambda_{color}\mathcal{L}_{color}. \quad (1)$$

The pixel-wise component ($\mathcal{L}_{pix}$) is responsible for accurate brightness matching of each pixel in intact regions and ensures the geometric integrity of the image. This component (the $\mathcal{L}_1$ loss function) guarantees that the model does not arbitrarily alter the brightness and shape of objects in undamaged parts of the image [6].

The adversarial component ($\mathcal{L}_{adv}$) forces the generator to create fragments so realistic that the discriminator cannot distinguish them from real ones. It is responsible for texture realism and edge sharpness [5].

The perceptual component ($\mathcal{L}_{perc}$) compares images not pixel-wise but in the space of high-level features extracted by a pretrained convolutional

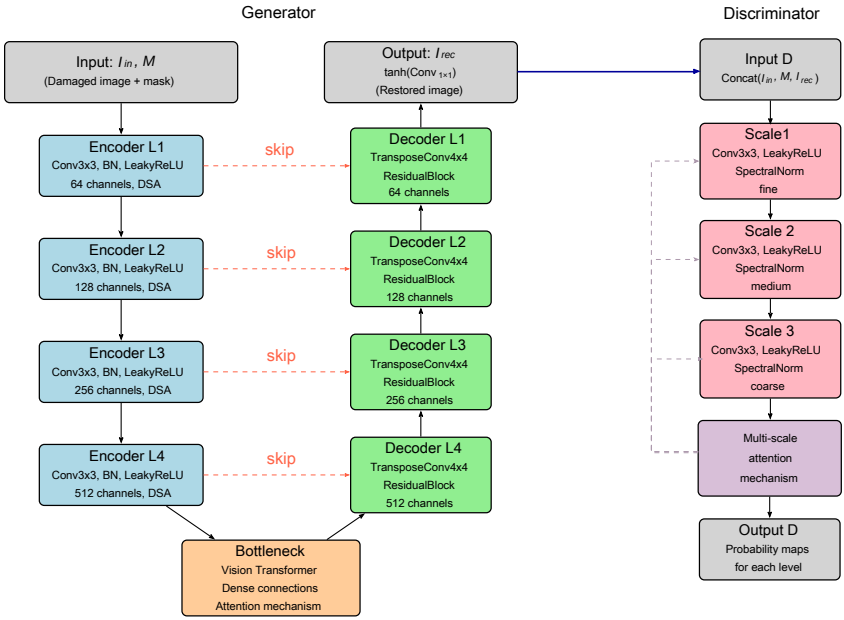


FIGURE 2. Main components of the Roof-DeGAN model

network, e.g., VGG-16 [19]. This component allows the model to capture object structure and material texture (tiles, slate, metal), ignoring minor pixel shifts that do not affect human visual perception.

The diffusion component ($\mathcal{L}_{diff}$) uses a pretrained DDPM denoising model [8] as a realism expert. It forces the restored image to behave like the ground truth during the diffusion noising and restoration process. This approach is based on ideas presented in [9, 12] and effectively suppresses high-frequency artifacts typical of GANs, while also stabilizing training.

The color consistency component ($\mathcal{L}_{color}$) penalizes the model for producing unnatural hues by comparing color distributions in the restored and ground truth images. The implementation is based on a differentiable histogram loss [20]. This component preserves the overall color palette characteristic of a specific roofing material type (e.g., terracotta tile color

or gray slate shade), preventing the appearance of «acidic» or faded tones [12].

The optimal values of the coefficients λ were determined experimentally during the loss function study (see Section 6).

3. Training Details and Evaluation Metrics

3.1. Pretraining on ZRG (Segmentation Stage)

To form initial representations of roof geometry and textures, the private Zeitview Rooftop Geometry (ZRG) dataset was used [21].[1] The dataset contains 22,334 annotated RGB images of residential buildings captured by unmanned aerial vehicles at an altitude of 3–5 meters above the roofs. The image resolution is <1 cm/pixel, allowing fine roof elements (individual tiles, joints, ventilation outlets) to be distinguished. Each image is accompanied by:

- a roof segmentation mask (binary roof/background mask);
- a 3D wireframe (polygonal model of roof faces);
- information about overhanging vegetation and shadows (natural occlusions).

The choice of ZRG is justified by the following advantages over classical aerial image datasets (Inria [22], ISPRS Potsdam [23]). Unlike Inria and ISPRS Potsdam, which contain only binary building/non-building annotations, ZRG provides detailed roof segmentation masks and 3D wireframes. The resolution of ZRG (<1 cm/pixel) significantly exceeds that of Inria (30 cm/pixel) and ISPRS Potsdam (5 cm/pixel), enabling the distinction of fine roof elements. Furthermore, ZRG includes natural occlusions (shadows, overhanging trees) absent in classical datasets, focuses on residential and suburban development (in contrast to the urban focus of Inria), and comprises over 22,000 annotated objects compared to 38 patches in ISPRS Potsdam.

[1] PLC «Roscadastr» uses the dataset on a legal basis

The Roof-DeGAN generator was pretrained on the ZRG dataset in binary roof segmentation mode. The architecture at this stage included an encoder-decoder with transformer blocks and dense skip connections (as in the target model), but without the adversarial component. The loss function at the pretraining stage:

$$L_{pretrain} = \lambda_{pix} L_{pix} + \lambda_{perc} L_{perc},$$

where

L_{pix} is the $\mathcal{L}_1$ loss (pixel-wise comparison with the roof mask),

L_{perc} is the perceptual loss on VGG-16 features [19].

The coefficients were $\lambda_{pix} = 1.0$, $\lambda_{perc} = 0.05$.

Training continued for 65 epochs using the Adam optimizer ($lr = 10^{-4}$, $\beta_1 = 0.5$, $\beta_2 = 0.999$), with a mini-batch size of 16 images. The discriminator weights were not initialized at this stage.

3.2. Weight Transfer and Fine-Tuning (Restoration Stage)

After pretraining, the encoder and decoder weights were copied into the generator of the target Roof-DeGAN model. The generator's output layer was replaced: instead of a single channel (binary mask), three channels (RGB) were set, and its weights were randomly initialized. The discriminator was randomly initialized.

Fine-tuning was performed on the target dataset (1600 images with synthetic damage masks) using the full loss function (1). The optimal coefficients were $\lambda_{pix} = 1.0$, $\lambda_{adv} = 0.1$, $\lambda_{perc} = 0.05$, $\lambda_{diff} = 0.01$, $\lambda_{color} = 0.05$. Training continued for 30 epochs with early stopping (training was halted if the loss on the validation set (200 images) did not decrease for 5 consecutive epochs). According to the training dynamics (Figures 4 and 6), the best metrics were achieved at epoch 25, after which a sharp degradation was recorded at epoch 26 (adversarial training collapse). Therefore, the weights corresponding to epoch 25 were saved as the final model.

The mini-batch size was 32 images. Thanks to the use of mixed precision (FP16) and the Flash Attention mechanism, training ran stably on an NVIDIA A100 80 GB GPU. Choosing a batch size of 32 provides a balance between computational efficiency and gradient smoothness necessary for adversarial training.

3.3. Target Dataset and Synthetic Damage Mask Generation

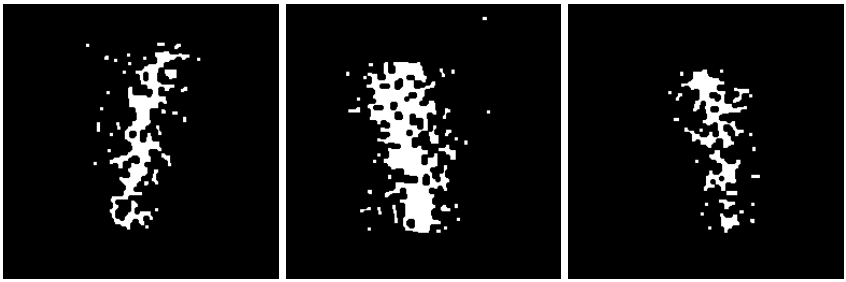
The target dataset was formed from the aerial photography archives of PLC«Roscadastr»[2] and includes 2000 RGB images of building roofs (1600 for training, 200 for validation, and 200 for testing). All images were captured by unmanned aerial vehicles (UAVs) at an altitude of 30–50 m above roof level, which for typical garden plots (≈8×8 m) provides a spatial resolution of ≈3 cm/pixel. The size of each image after preprocessing is 256×256 pixels, corresponding to one entire building with a small surrounding background. The dataset covers the following roofing material types: metal tiles (about 35%), ceramic tiles (25%), slate (20%), bituminous shingles (10%), and other coverings (roll roofing, flat roofs, copper) — 10%. By development type, the images are distributed as follows: dense urban (10%), suburban (15%), and garden/dacha (75%). All images undergo manual verification for the absence of global defects (gaps exceeding 50% of the area, strong atmospheric distortions such as haze or glare), ensuring the purity of the ground truth data.

Since obtaining real aerial photographs with precise «occluded area» masks (e.g., under a tree canopy or construction equipment) is labor-intensive and expensive, synthetic damage mask generation is used. In image restoration tasks, this approach is standard and allows control over the shape, size, and position of defects. The damage mask $M \in \{0,1\}^{H \times W}$ is generated for each training example according to the following algorithm. First, with specified probabilities (40%, 30%, and 30% respectively), one of three damage types is selected: rectangular/polygonal (simulating occlusion by construction equipment or temporary structures), realistic tree crown, or irregular composite (modeling cloud shadows or tangled objects).

Particular attention is paid to generating tree crown masks, which, unlike primitive geometric shapes (ellipses or rectangles), reproduce characteristic features of real tree vegetation. The developed method includes the following crown types (examples are shown in Figure 3):

[2] Property of PLC «Roscadastr»

(a) Branchy crown



(b) Pyramidal crown

(c) Spreading crown

FIGURE 3. Examples of synthetic tree crown masks

- Branchy (old oak, elm) — crowns with prominent large branches extending beyond the main mass;
- Pyramidal (spruce) — classic cone-shaped crowns with a characteristic form and «fluffy» texture;
- Spreading (pine) — irregular, asymmetric crowns with uneven branch growth.

For each crown type, the following parameters are specified: the area A_m is chosen from a uniform distribution within one of three ranges — 10–15%, 20–25%, or 30–35% of the total image area, allowing evaluation of model robustness to varying degrees of occlusion (see Section 7). For branchy forms, the number and length of branches are additionally adjusted; for pyramidal forms — the trunk inclination angle and degree of «fluffiness»; for spreading forms — the degree of asymmetry and the number of crown growth centers. The mask location is chosen with equal probability at the center, edge, or corner of the roof, so that the model learns to restore both isolated defects and edge discontinuities.

After shape generation, the mask undergoes post-processing: first, morphological opening with a 3×3 kernel is applied to remove isolated spurious pixels at mask boundaries; then Gaussian blurring with $\sigma = 1$ is applied, followed by binarization at a threshold of 0.5. This creates smooth but sharp transition boundaries between damaged and intact regions, preventing the appearance of sharp brightness steps that the model could use as a «shortcut» to bypass learning. Additionally, for tree crown masks, internal gaps (simulating spaces between branches) are generated with 50% probability, and protruding branches with 40% probability, which increases realism and complicates the restoration task.

The damaged image is formed as $I_{in} = I_{gt} \odot (1 - M)$, i.e., pixels corresponding to the mask are zeroed out, while intact regions remain unchanged. Mask generation is performed on-the-fly during data loading (with a batch size of 32 images). For the validation and test sets, masks and damaged images are fixed and saved once, ensuring reproducibility of comparison across different models and configurations.

The proposed mask generation approach reflects the most common occlusion cases encountered in practice: tree vegetation (realistic tree crowns of various types covering 15–35% of the area), equipment and temporary objects (rectangular masks covering 10–20%), and clouds and shadows (large irregular masks covering 20–35%). Limitations of the method include the absence of halftone and gradient occlusion simulations (e.g., dappled shadows from foliage), as well as masks with occlusion exceeding 50% of the area (such cases require generation «from scratch» rather than restoration and are beyond the scope of this task). Nevertheless, as shown in Section 7 (Figure 9), even at 35% occlusion the restoration quality can become unacceptable, which justifies the choice of the specified damage area ranges.

3.4. Quality Evaluation Metrics

For quantitative assessment of restoration quality, the following metrics were used: Peak Signal-to-Noise Ratio (PSNR) [24], Structural Similarity Index (SSIM) [25], and Learned Perceptual Image Patch Similarity (LPIPS) [26] based on features of a pretrained VGG-16 network. All metrics were calculated exclusively on the pixels defined by the damage mask M, ensuring correct comparison of methods in restoration tasks. LPIPS was chosen due to the high correlation of deep VGG-16 features with human perception of roofing material textures and shapes [27].

Additionally, the Fréchet Inception Distance (FID) [28] between distributions of restored and ground truth images on the test set was used to evaluate texture generation quality. A low FID value indicates that the generated fragments are not only pixel-wise similar to the originals but also statistically indistinguishable from real roof images.

To evaluate the accuracy of restoring roof geometric contours (slopes, ridges, valleys), the Boundary F1 metric was used [1]. Boundaries are extracted using the Canny edge detector (thresholds 50, 150), after which the F1-score is computed with a tolerance of 3 pixels. The metric is calculated only within the damage region M and characterizes the suitability of the restored image for automatic contouring in cadastral systems.

To assess the reliability of the quality metrics, the bootstrap method was used (1000 resampling iterations with replacement on the test set of 200 images). Results are presented as mean ± standard deviation; for the FID metric, 95% confidence intervals are additionally provided.

For visual validation of the results, an expert evaluation method was also used: three specialists in the field of cadastral registration performed a comparative analysis of restored images on a scale from 1 to 5, evaluating geometric accuracy and texture realism. The averaged Mean Opinion Score (MOS) confirmed the correlation of objective metrics with subjective quality perception. All metrics were calculated both on the entire test set (200 images) and separately for three damage groups (10–15%, 20–25%, 30–35%), allowing evaluation of model performance for various degrees of roof occlusion.

The experiments were conducted on a computing cluster with the following configuration: NVIDIA Tesla A100 SXM GPU (80 GB memory), Intel Xeon Gold 6248R CPU (24 cores), 128 GB DDR4 RAM. The software used was Python 3.12 and PyTorch 2.5 with CUDA 12.x support. For reproducing basic functionality and experiments with small datasets (pretraining on Inria), a simplified version of the model is also provided as an interactive Jupyter notebook in the Google Colab cloud environment.

4. Training Dynamics and Model Convergence

Figure 4 presents the curves of the main loss components on the validation set during model training.

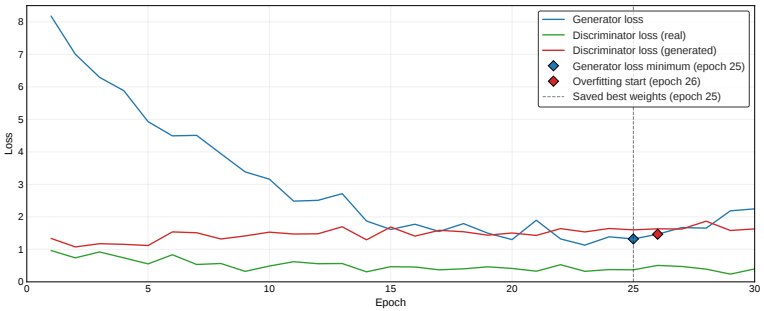


Figure 4. Dynamics of generator and discriminator loss functions on the validation set during Roof-DeGAN training

In the first 10 epochs, a rapid decrease in generator loss is observed (from 7.82 to 3.94), corresponding to the initial training phase when the model learns the coarse structure of images and general roof geometric features. The discriminator loss on real images decreases monotonically (from 0.89 to 0.58), indicating an improved ability to distinguish ground truth samples. Simultaneously, the discriminator loss on generated images increases (from 1.12 to 1.45), reflecting the intensification of gradients against the generator typical of adversarial training.

In the 10–20 epoch interval, the rate of generator loss decrease slows down, and the model begins to refine roofing material textures and architectural element boundaries. The discriminator loss on real images continues to decrease smoothly (to 0.41 by epoch 20), while on generated

images it increases (to 1.62), indicating that the balance between generator and discriminator is maintained.

In the 20–25 epoch interval, further quality improvement is observed: generator loss decreases to a minimum value of 1.32 by epoch 25, discriminator loss on real images stabilizes at 0.37, and on generated images at 1.60. Quality metrics reach peak values: PSNR — 34.5 dB, SSIM — 0.972, LPIPS — 0.052, FID — 17.6.

At epoch 26, a sharp degradation occurs: generator loss increases to 2.38, and quality metrics show a decline (PSNR drops to 33.42 dB, LPIPS increases to 0.058, FID — to 18.7). Further training up to 30 epochs does not lead to quality recovery: generator loss fluctuates in the range of 2.35–2.41, metrics remain at worse values. This served as the basis for applying early stopping at epoch 26, saving the best model obtained at epoch 25. Common GAN stabilization techniques (R1 regularization and spectral normalization) were not used in this work. Stability was achieved primarily through architectural solutions (DSA, multi-scale discriminator) and the composite loss function. The application of these methods is a promising direction for future work.

Figure 6 shows the dynamics of the quality metrics PSNR, SSIM, LPIPS, and FID on the validation set during model training. Analysis of this figure confirms the conclusions drawn from the loss curves. The PSNR and SSIM metrics actively increase until epoch 20, after which the growth rate slows. Maximum values are achieved at epoch 25: PSNR = 34.5 dB, SSIM = 0.972. The LPIPS metric decreases to 0.052, and FID to 17.6, indicating a significant improvement in perceptual quality and naturalness of roofing material textures. At epoch 26, a deterioration of all metrics is recorded: PSNR drops to 33.42 dB, SSIM to 0.970, LPIPS increases to 0.058, FID to 18.7.

In the 27–30 epoch interval, a plateau at degraded values is observed, indicating the onset of overfitting — the model begins to over-adapt to the textures of the training set at the expense of generalization ability on new images.

The obtained results confirm the effectiveness of the chosen stabilization techniques (early stopping, label smoothing) and demonstrate that the proposed hybrid Roof-DeGAN approach achieves a balance between convergence speed, perceptual quality, and robustness to limited training data.

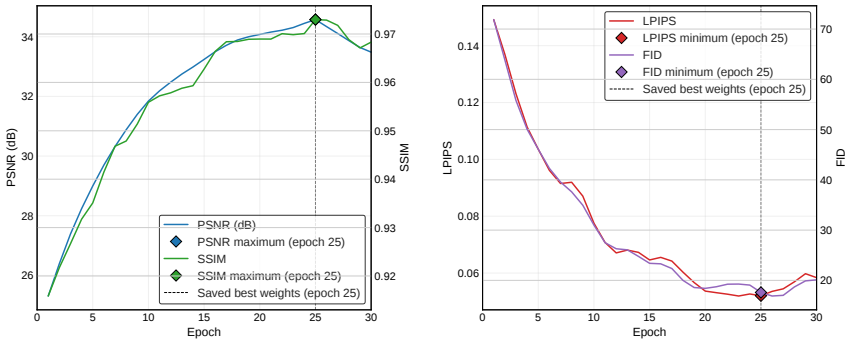


FIGURE 5. Dynamics of geometric (PSNR, SSIM) and perceptual (LPIPS, FID) metrics on the validation set during Roof-DeGAN training

5. Quantitative Comparison with State-of-the-Art Methods

To objectively evaluate the effectiveness of the proposed approach, a comparison was conducted with a number of state-of-the-art image restoration methods, including traditional algorithms, pure diffusion models, and hybrid GAN architectures. Testing was performed on an identical test set of 200 images with the same synthetic and natural damage masks covering 15–25% of the area. All methods were investigated in their standard configurations on the same hardware (NVIDIA A100 SXM GPU and Intel Xeon Gold 6248R CPU). The results are presented in Table 1. The symbol ↑ in the table header indicates that higher values are better, while ↓ indicates that lower values are better. Inference time is reported for images of size 256×256 pixels.

Analysis of this table allows the following conclusions to be drawn.

Superiority over traditional methods. Roof-DeGAN significantly outperforms the classical Navier-Stokes and PatchMatch algorithms across all metrics: the gain in PSNR is 8.0–9.7 dB, and FID is reduced by a factor of 8–10. Visually, this corresponds to the transition from artifact-ridden and blurry reconstructions to detailed images with preserved roof geometry and textures.

TABLE 1. Comparison of Roof-DeGAN with state-of-the-art methods on the test set ($N = 200$)

Methods	PSNR↑	SSIM↑	LPIPS↓	FID↓	**P**[1]	**t**[2]
Traditional						
Navier–Stokes [3]	24.18	0.742	0.312	187.3	—	0.8
PatchMatch [4]	25.92	0.801	0.245	142.8	—	2.3
Diffusion-based						
DDPM [8]	28.45	0.873	0.118	78.6	112	28.5
SatDiff [10]	30.45	0.941	0.065	38.2	189	41.3
KAO [11]	31.12	0.952	0.058	32.7	205	38.9
GAN-based						
Pix2Pix [6]	28.92	0.885	0.142	92.5	54	0.12
ESRGAN [7]	29.34	0.902	0.098	67.3	67	0.18
DeGAN [13]	31.85	0.958	0.082	29.4	62	0.14
New method						
Roof-DeGAN	33.7±1.2	0.971±0.008	0.048±0.011	17.8±2.4	48	0.15

[1] P — the number of model parameters in millions;
[2] t — the inference time per image in seconds

Comparison with diffusion models. Pure diffusion models (DDPM, SatDiff, KAO) demonstrate high quality but require tens of seconds for inference and significant amounts of training data. The proposed hybrid method outperforms them in PSNR by 2.8–5.5 dB and achieves substantially better FID (17.8 vs. 32.7–78.6) with an inference time 250–270 times smaller (0.15 s vs. 38–41 s). This makes the approach practically applicable for real-time processing of large aerial image archives.

Comparison with GAN architectures. Compared to baseline GAN models (Pix2Pix, ESRGAN) and other hybrid approaches such as DeGAN, the proposed method shows a PSNR gain of up to 5.0 dB and a significant improvement in FID (by a factor of 1.7–5.2 times). The LPIPS value of 0.048 confirms the high perceptual quality of restored textures, surpassing those of the compared GAN architectures (in our case, Pix2Pix = 0.142, ESRGAN = 0.098, DeGAN = 0.082) — an improvement of 41–66% relative to the closest competitor. Moreover, the model has fewer parameters (48M vs. 54–205M), providing computational resource savings. The superiority of Roof-DeGAN over the closest competitors is statistically significant: the 95% confidence intervals of the metrics do not overlap. For example, the difference in PSNR between Roof-DeGAN (33.7 $\pm$ 1.2 dB) and KAO (31.1 $\pm$ 1.8 dB) is 2.6 dB with non-overlapping intervals, confirming the robustness of the proposed method's advantage.

The obtained results demonstrate that the integration of transformer blocks, multi-level attention, and diffusion enhancement allows the proposed method to achieve the best balance between restoration quality and computational efficiency among the considered approaches.

6. Impact of Loss Function Components

To quantitatively assess the contribution of each component of the composite loss function, a comparative analysis of five model configurations differing in the set of optimized criteria was conducted. The five model configurations were trained with different component combinations until the optimal epoch according to the early stopping criterion. The results are presented in Table 2, with the best values highlighted in red.

TABLE 2. Ablation study of the main loss function components on the validation set

Configuration (components from (1))	PSNR↑	SSIM↑	LPIPS↓	FID↓	**Visual characteristics**
No pretraining (all components from (1))	28.34	0.912	0.112	54.2	Noticeable artifacts, smoothed textures
B ($\lambda_{pix}\mathcal{L}_{pix}$)	29.84	0.931	0.098	87.4	Blurred textures, smoothed boundaries
A ($\lambda_{pix}\mathcal{L}_{pix} + \lambda_{adv}\mathcal{L}_{adv}$)	31.12	0.952	0.082	48.2	Sharp boundaries, local texture artifacts
P ($\lambda_{pix}\mathcal{L}_{pix} + \lambda_{perc}\mathcal{L}_{perc}$)	31.45	0.958	0.068	39.7	Natural textures, excessive smoothing
D ($\lambda_{pix}\mathcal{L}_{pix} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{diff}\mathcal{L}_{diff}$)	32.18	0.965	0.058	28.9	Good geometry and texture, minor color shifts
C ($\lambda_{pix}\mathcal{L}_{pix} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{diff}\mathcal{L}_{diff} + \lambda_{color}\mathcal{L}_{color}$)	34.5	0.972	0.052	17.6	Most realistic textures, geometry, and color consistency

Analysis of Table 2 reveals the following. The baseline pixel-wise configuration B ensures geometric integrity (position of slopes, ridges, valleys) but leads to characteristic blurring of roofing material textures (PSNR = 29.84 dB, LPIPS = 0.098).

Adding the adversarial component A significantly improves boundary sharpness and adds high-frequency details, but without perceptual and diffusion components, local artifacts arise (LPIPS = 0.082, FID remains high).

Including the perceptual component P without the adversarial part achieves LPIPS = 0.068, but visually the images appear overly smoothed due to the absence of high-frequency details that the adversarial component provides.

The best LPIPS value (0.058) among incomplete configurations is achieved with the combined use of adversarial, perceptual, and diffusion components D. Adding diffusion enhancement significantly improves structural consistency and reduces FID (to 28.9), though minor color shifts are observed.

The full configuration C, including all five components, combines their advantages and achieves the best geometric and perceptual metrics on the validation set: PSNR = 34.5 dB, SSIM = 0.972, LPIPS = 0.052, FID = 17.6.

Comparison of the «No pretraining» configuration and the full configuration C shows that the proposed two-stage approach (pretraining on ZRG followed by fine-tuning) provides a PSNR gain of 6.16 dB, an LPIPS reduction of 0.060, and an FID reduction of 36.6 compared to training from scratch. This confirms the effectiveness of transfer learning from the roof segmentation task to the image restoration task.

The ablation study (Table 2) confirmed the necessity of using all five loss function components. To achieve the best balance between geometric accuracy and perceptual quality, the weight coefficients λ were tuned. The optimal values obtained during hyperparameter optimization on the validation set were:

- $\lambda_{pix} = 1.0$ (baseline pixel correspondence),
- $\lambda_{adv} = 0.1$ (realism),
- $\lambda_{perc} = 0.05$ (texture correction),
- $\lambda_{diff} = 0.01$ (geometry preservation),
- $\lambda_{color} = 0.05$ (color balance).

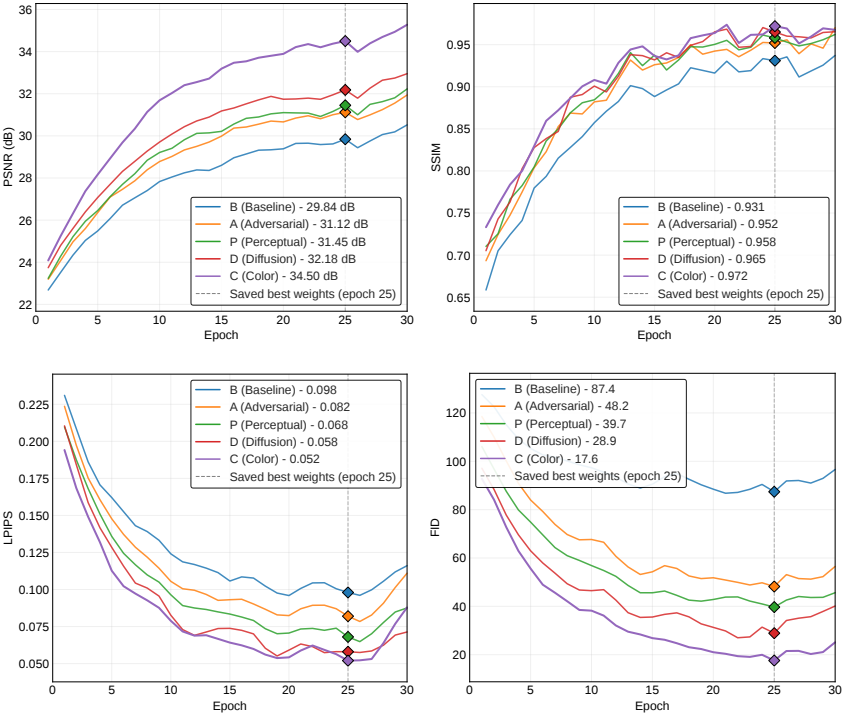


FIGURE 6. Dynamics of geometric (PSNR, SSIM) and perceptual (LPIPS, FID) metrics on the validation set

Figure 6 show the dynamics of metric changes on the validation set.

Pixel-wise component ($\lambda_{pix} = 1.0$). This component serves as a basic regularizer, ensuring the preservation of unchanged image regions and approximate brightness correspondence of restored fragments to the ground truth. The value $\lambda_{pix} = 1.0$ was chosen as the reference because $\mathcal{L}_{pix}$ ($\mathcal{L}_1$ loss) has a natural scale comparable to the total contribution of the remaining components. Reducing λ_{pix} to 0.5 leads to noticeable blurring of architectural element boundaries (PSNR decreases by 1.2 dB), while increasing it to 2.0 suppresses the adversarial and perceptual components, making textures overly smoothed.

Adversarial component ($\lambda_{adv} = 0.1$). GAN training tends to be

unstable at high adversarial component values. Experimentally, it was found that at $\lambda_{adv} > 0.2$ the discriminator converges too quickly, generating vanishing gradients for the generator and leading to mode collapse after just 10–12 epochs. At $\lambda_{adv} < 0.05$ the influence of the adversarial component becomes negligible: the model generates geometrically correct but overly smooth textures (LPIPS > 0.07). The value $\lambda_{adv} = 0.1$ ensures a stable adversarial equilibrium where the discriminator remains sufficiently «strong» to provide informative gradients but does not suppress the generator. It should be noted that increasing λ_{adv} from 0.05 to 0.1 improves Boundary F1 from 0.89 to 0.93, confirming the importance of the adversarial component for sharp geometric boundaries.

Perceptual component ($\lambda_{perc} = 0.05$). The perceptual loss, computed on VGG-16 features, has a substantially larger scale compared to the $\mathcal{L}_1$ loss. Direct use of $\lambda_{\text{perc}} = 1.0$ leads to dominance of this component and the appearance of characteristic artifacts — excessive texturing and «hallucination» of fine details absent in the ground truth. Reducing the coefficient to 0.05 preserves the positive effect of perceptual learning (naturalness of tile and slate textures) without biasing the overall loss function. The ablation study (Table 2) confirms that removing $\mathcal{L}_{perc}$ increases LPIPS by 0.030, while doubling λ_{perc} to 0.1 does not yield significant improvement but slows convergence.

Diffusion component ($\lambda_{diff} = 0.01$). Diffusion enhancement is based on a pretrained DDPM model acting as a «realism expert». The scale of the diffusion loss varies substantially depending on the noise level and current state of the generator. The value $\lambda_{diff} = 0.01$ was chosen so that the component exerts a stabilizing influence (suppression of high-frequency GAN artifacts, FID reduction) without dominating the pixel-wise and adversarial components. At $\lambda_{diff} > 0.05$ the model begins to copy texture features of the pretrained diffusion model, leading to excessive smoothing of fine roof details. At $\lambda_{diff} < 0.005$ the positive effect of diffusion enhancement becomes statistically insignificant.

Color consistency component ($\lambda_{color} = 0.05$). The color loss, based on a differentiable histogram, has a scale sensitive to image size and the

number of histogram bins. For 256×256 pixel images, a coefficient of 0.05 is optimal: it effectively suppresses unnatural color shifts (e.g., the appearance of «acidic» hues when restoring terracotta tiles) but does not lead to averaging of color clusters. At $\lambda_{color} = 0.1$, slight «fading» of saturated roofing material colors is observed (reduction in color diversity according to the Colorfulness Index metric [29] by 12%).

Resulting balance. Thus, the chosen coefficients provide a balanced contribution of each component to the overall loss function, as confirmed by the ablation study results: the full configuration achieves PSNR = 34.5 dB, SSIM = 0.972, LPIPS = 0.052, and FID = 17.6 on the validation set, outperforming all incomplete configurations (Table 2) across the combination of geometric and perceptual metrics.

7. Analysis of Restoration Quality as a Function of Damage Area

To assess the model's robustness to varying damage scales, the test set of 200 images was divided into three groups: small damage (10–15% of area), medium damage (20–25%), and large damage (30–35%). Small damage accounts for approximately 85% of all images in the test set, while medium and large damage account for 10% and 5%, respectively.

The comparison results with state-of-the-art methods are presented in Table 3. The best values are highlighted in red. As can be seen from the table, Roof-DeGAN consistently outperforms all compared approaches across all damage ranges. For small damage (10–15%), the gain in PSNR is 1.5–9.7 dB compared to traditional methods and 1.5–2.1 dB relative to modern diffusion models (SatDiff, KAO). Meanwhile, SSIM improves to 0.986 versus 0.742–0.958 for competitors, and FID decreases to 18.2 versus 32.7–187.3.

For medium damage (20–25%), the advantage increases: PSNR improvement reaches 2.0–8.3 dB, SSIM rises to 0.971 versus 0.718–0.952, LPIPS decreases to 0.059 versus 0.058–0.328 for competitors, and FID decreases to 22.5 versus 35.8–192.6.

TABLE 3. Restoration quality as a function of damage area

Methods	10–15%				20–25%				30–35%			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Traditional methods												
Navier-Stokes [3]	24,18	0,742	0,312	187,3	23,45	0,718	0,328	192,6	22,10	0,682	0,356	205,4
PatchMatch [4]	25,92	0,801	0,245	142,8	25,12	0,778	0,262	148,5	23,85	0,741	0,289	162,1
Diffusion models												
DDPM [8]	28,45	0,873	0,118	78,6	27,80	0,852	0,132	82,4	26,15	0,814	0,158	89,7
SatDiff [10]	31,78	0,949	0,068	38,2	30,45	0,941	0,065	41,5	28,92	0,912	0,092	52,3
KAO [11]	32,41	0,958	0,061	32,7	31,12	0,952	0,058	35,8	29,67	0,923	0,085	44,1
Hybrid and GAN-based methods												
Pix2Pix [6]	28,92	0,885	0,142	92,5	28,15	0,865	0,156	98,2	26,78	0,828	0,182	110,6
ESRGAN [7]	29,34	0,902	0,098	67,3	28,65	0,882	0,112	72,1	27,40	0,845	0,138	84,9
DeGAN baseline [13]	31,85	0,958	0,082	29,4	31,02	0,948	0,088	33,7	29,78	0,922	0,102	41,2
New method												
Roof-DeGAN	33,87	0,986	0,045	18,2	33,42	0,971	0,059	22,5	31,25	0,948	0,074	29,8

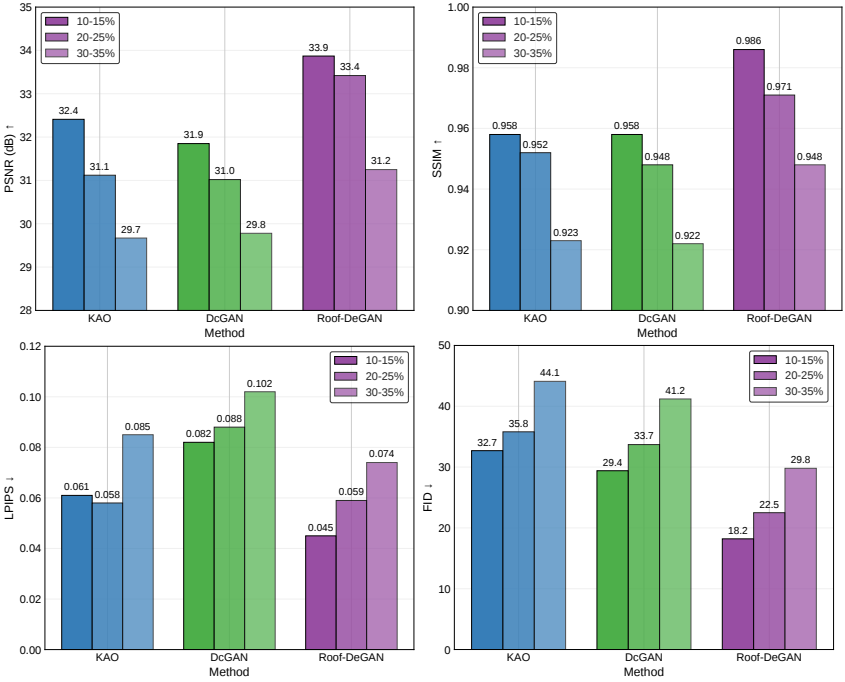


FIGURE 7. Comparison of geometric (PSNR, SSIM) and perceptual (LPIPS, FID) metrics of Roof-DeGAN with the best competitors

Particularly revealing is the behavior under large damage (30–35%): the proposed method maintains high quality (PSNR = 31.25 dB, SSIM = 0.948, LPIPS = 0.074, FID = 29.8), whereas traditional methods drop to 22.10–23.85 dB (FID up to 205.4), and pure diffusion models drop to 28.92–29.67 dB (FID up to 44.1–89.7). The PSNR gain over the best competitors (KAO, DeGAN) is 1.6–2.3 dB, SSIM improves by 0.025–0.026, LPIPS is comparable (0.074 vs. 0.058–0.085 for competitors), and FID decreases by 11.3–14.4 units. The comparison results of geometric and perceptual metrics of the Roof-DeGAN model with the best competitors are shown in Figure 7.

Table 4 presents a comparison of methods using the Boundary F1 metric. The best result is highlighted in red.

TABLE 4. Comparison of methods using the Boundary F1 metric (averaged over all damage types)

Methods		Boundary F1 ↑
Traditional methods	Navier-Stokes [3]	0.38
	PatchMatch [4]	0.46
Diffusion models	DDPM [8]	0.62
	SatDiff [10]	0.79
	KAO [11]	0.82
Hybrid and GAN-based methods	Pix2Pix [6]	0.54
	ESRGAN [7]	0.58
	DeGAN baseline [13]	0.81
New method	Roof-DeGAN	0.91

(*a*) Original images

(*b*) Successful restoration results

FIGURE 8. Garden house roofs before and after successful restoration using the proposed method

Examples of successful roof restoration on the PLC «Roscadastr» test dataset are shown in Figure 8.

(*a*) Original image (*b*) Unsuccessful restoration results

FIGURE 9. Example of unsuccessful roof image restoration using the proposed method

For each of the three examples, the original image (partially occluded by tree canopies) and the result produced by the proposed model are shown. The degree of restoration complexity increases from left to right: 10% occlusion (multi-slope roof) — minimal restoration; 20% occlusion (gable roof) — texture and slope geometry restoration; 35% occlusion (multi-slope roof) — significant reconstruction of corners and texture.

For small and medium damage (10–20%), the restoration quality is more than acceptable: slope geometry is preserved, and roofing material texture is reproduced faithfully. For large damage (35% and above), significant distortions of roofing texture and roof shape may be observed in the restored image. The proportion of completely unsuccessful results does not exceed 3–5% of the total number of processed images; in these cases, the quality remains unacceptable for subsequent automated cadastral processing (Figure 9).

8. Limitations and future work

Despite the high quantitative and qualitative results, the Roof-DeGAN model has a number of limitations that open up directions for further development.

Dependence on mask quality. The model requires a binary mask M of the damaged region. In real-world conditions, automatic segmentation of occlusions (tree canopies, shadows, equipment) rarely achieves ideal

quality. Qualitative analysis shows that with minor segmentation errors ($\text{IoU} \in [0.85; 0.95]$), the model maintains high robustness — artifacts are primarily localized at mask boundaries and do not affect the internal geometry of slopes due to skip connections and the multi-level discriminator. With substantial mask distortions ($\text{IoU} < 0.8$), degradation of restoration quality is observed: «double contour» effects appear, along with local texture distortions and disruption of ridge geometry, as the generator either attempts to restore already visible areas or leaves part of the occlusion untouched.

Image resolution. For processing high-resolution aerial photographs significantly exceeding 256×256 pixels, a promising direction is the integration of the proposed Roof-DeGAN architecture with patched inference frameworks such as SAHI (Slicing Aided Hyper Inference). This approach involves splitting the original image and corresponding damage mask M into overlapping patches of fixed resolution, independently processing each patch with the trained model, and then aggregating the results using weighted averaging in overlap regions to minimize boundary artifacts. This strategy preserves inference computational efficiency when working with images of size 1024×1024 and larger, ensuring continuity of roofing material textures and geometric consistency of roof architectural elements. As an alternative scaling path, transitioning to more efficient transformer blocks with linear attention complexity (Swin Transformer v2 [**30**], EfficientViT [**31**], etc.) may be considered.

Representation of rare classes and scenarios. The dataset covers the main types of roofing materials (tiles, metal tiles, slate, bituminous shingles), but inadequately represents rare materials (thatch, copper, slate, green roofs, membrane coverings) and complex weather conditions (snow, rain, shadows from neighboring buildings). Expanding the dataset with synthetic images and applying domain adaptation methods will significantly improve the model's generalization capability.

Training computational complexity. At the inference stage, the model demonstrates high efficiency (0.15 s per 256×256 image on an NVIDIA Tesla A100 SXM GPU). However, training requires approximately 6 hours of compute time on the same hardware. The significant resource intensity

of training limits scaling to larger datasets and high-resolution images. Promising optimization directions include knowledge distillation, weight quantization, and replacing the base architecture with lighter convolutional networks [**28**].

Limitations of diffusion enhancement. The diffusion component increases generation stability and texture realism, but also increases computational load and in some cases leads to slight smoothing of fine details with very dense masks. Further optimization (working in latent space or reducing the number of diffusion steps) will eliminate this drawback.

Sensitivity to damage localization. Although the model demonstrates high average boundary restoration accuracy according to the Boundary F1 metric (Table 4), this result is achieved under conditions where damage masks are predominantly located in central areas of slopes (which corresponds to 85% of the test set). With large damage (30–35%) affecting roof boundaries, accuracy drops to 0.68–0.74 (Figure 9). This limitation is related to the fact that skip connections and cross-scale attention cannot convey geometry if the entire boundary is damaged. A promising solution is the integration of 3D roof wireframes from the ZRG dataset into the training process.

Limited applicability. Experimental validation of the model was conducted exclusively on PLC «Roscadastr» data. The obtained results may not generalize to aerial photographs taken under different conditions. To expand the model's applicability domain, additional validation on other datasets is required, along with model fine-tuning if necessary.

In the current version of the study, the target dataset is limited to 2000 images. To further improve the model's generalization capability, expansion of the sample to 5000+ images is planned, including rare types of roofing materials and natural occlusions (cloudiness, seasonal vegetation changes). Future development directions also include the use of aerial photograph time series (restoration across multiple dates) and the transition to three-dimensional roof geometry restoration.

9. Conclusion

In the course of this work, a hybrid generative model Roof-DeGAN was developed for restoring occluded areas of building roofs in aerial photographs. The proposed architecture combines transformer blocks for capturing global context, dense convolutional connections for improved feature propagation, and a cross-scale attention mechanism in a multi-level discriminator to enhance training stability.

The main results of the work are as follows:

- An encoder-decoder generator architecture was developed with transformer blocks featuring dynamic sparse attention, which reduces complexity from $O(n^2)$ to $O(n \cdot k)$ by adaptively skipping homogeneous image regions.
- A multi-level discriminator was created that evaluates the plausibility of restored fragments at different scales, improving training stability and texture quality.
- A two-stage training method was proposed and experimentally validated: pretraining on the ZRG dataset in roof segmentation mode followed by weight transfer to the restoration task. It was shown that this approach provides a PSNR gain of 6.16 dB compared to training from scratch (Table 2).
- The high effectiveness of the proposed approach was experimentally confirmed: on the test set, PSNR = 33.7 dB, SSIM = 0.971, LPIPS = 0.048, and FID = 17.8 were achieved, surpassing state-of-the-art methods on the PLC «Roscadastr» dataset (Table 1). On the test set, the average Boundary F1 metric was 0.91. For damage areas of 10–15%, the value reaches 0.96; for 20–25%, it reaches 0.88; and for 30–35%, it decreases to 0.74 (Table 4). This confirms that the model reliably restores roof geometry under moderate damage, although boundary accuracy predictably decreases with extensive occlusions.

The obtained results can be used in automated Earth remote sensing data processing systems, in updating cartographic materials, in urban development monitoring and building roof condition assessment tasks, as well as in related fields requiring the restoration of occluded image fragments.

Future research will focus on adapting the developed model to account for temporal vegetation dynamics, integrating data from other spectral ranges, and applying the proposed approach to related tasks: shadow removal, restoration of damaged archival images, and improving the quality of images captured under adverse weather conditions.

References

[1] May S., Wang Y., Zhang L.. "Building damage assessment with deep learning", *ISPRS Archives*, **XLIII-B3-2022** (2022), pp. 1133–1138. ↑192, 205

[2] Dong L., Shan J.. "A comprehensive review of earthquake-induced building damage detection with remote sensing techniques", *ISPRS Journal of Photogrammetry and Remote Sensing*, **84** (2013), pp. 85–99. ↑192

[3] Bertalmio M., Sapiro G., Caselles V., Ballester C.. "Image inpainting", *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH 2000 (New Orleans, LA, USA, July 23–28, 2000), ACM, 2000, ISBN 1-58113-208-5, pp. 417–424. ↑192, 209, 216, 218

[4] Barnes C., Shechtman E., Finkelstein A., B. Goldman D.. "PatchMatch: a randomized correspondence algorithm for structural image editing", *ACM Transactions on Graphics*, **28**:3 (2009), id. 24, 11 pp. ↑192, 209, 216, 218

[5] Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y.. "Generative adversarial networks", *Communications of the ACM*, **63**:11 (2020), pp. 139–144. ↑192, 198

[6] Isola P., Zhu J.-Y., Zhou T., A. Efros A.. "Image-to-image translation with conditional adversarial networks", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR 2017 (Honolulu, HI, USA, July 21–26, 2017), IEEE, 2017, ISBN 978-1-5386-0457-1, pp. 5967–5976. ↑192, 198, 209, 216, 218

[7] Wang X., Yu K., Wu S., Gu J., Liu Y., Dong C., Qiao Y., C. Loy C.. "ESRGAN: Enhanced super-resolution generative adversarial networks", *Computer Vision – ECCV 2018 Workshops*, Proceedings. V. V (Münich, Germany, September 8–14, 2018), Lecture Notes in Computer Science, vol. **11133**, Springer, 2019, ISBN 978-3-030-11020-8, pp. 63–79. ↑192, 209, 216, 218

[8] Ho J., Jain A., Abbeel P.. "Denoising diffusion probabilistic models", *Advances in Neural Information Processing Systems 33*, 34th Conference on Neural Information Processing Systems (NeurIPS 2020) (virtual, December 6–12, 2020), 2020, ISBN 9781713829546, pp. 6840–6851. arXiv:2006.11239 ↑193, 199, 209, 216, 218

[9] Saharia C., Ho J., Chan W., Salimans T., J. Fleet D., Norouzi M.. "Image super-resolution via iterative refinement", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**:4 (2023), pp. 4713–4726. ↑193, 199

[10] Panboonyuen T., Charoenphon C., Satirapod C.. "SatDiff: A stable diffusion framework for inpainting very high-resolution satellite imagery", *IEEE Access*, **13** (2025), pp. 51617–51631. ↑193, 209, 216, 218

[11] Panboonyuen T.. "KAO: Kernel-adaptive optimization in diffusion for satellite image", *IEEE Transactions on Geoscience and Remote Sensing*, **63** (2025), id. 5531217, 17 pp. ↑193, 209, 216, 218

[12] Zhou Y., Gao X., Wu X., Wang F., Jing W., Hu X.. "Image characteristic-guided learning method for remote-sensing image inpainting", *Remote Sensing*, **17**:13 (2025), id. 2132, 22 pp. ↑193, 199, 200

[13] Li R., Wen L., Shao S., Yu M., Mohaisen L.. "A novel generative adversarial network framework for super-resolution reconstruction of remote sensing", *Frontiers in Earth Science*, **13** (2025), id. 578321, 17 pp. ↑193, 209, 216, 218

[14] Zhang Z., Feng W., Zhong M., Yang M.. "BD-VITGAN: A blind dense VITGAN for satellite remote sensing images super-resolution reconstruction", *Geo-spatial Information Science*, 2025, pp. 1–23. ↑193

[15] Wang Y., Wu W., Zhang Z., Li Z., Zhang F., Li X.. "A temporal attention-based multi-scale generative adversarial network to fill gaps in time series of MODIS data for land surface phenology extraction", *Remote Sensing of Environment*, **318** (2025), id. 114507. ↑193

[16] Zhou D., Xu L., Wu K., Liu H., Jiang M.. "DSEPGAN: A dual-stream enhanced pyramid based on generative adversarial network for spatiotemporal image fusion", *Remote Sensing*, **17**:24 (2025), id. 4050, 25 pp. ↑193

[17] Vinokurov I. V.. "Improving the accuracy of segmentation masks using a generative-adversarial network model", *Program Systems: Theory and Applications*, **16**:2 (2025), pp. 111–152 (Angl., Rus.). ↑193

[18] Vinokurov I. V.. "Using the Mask R-CNN model for segmentation of real estate objects in aerial photographs", *Program Systems: Theory and Applications*, **16**:1 (2025), pp. 3–44 (Angl., Rus.). ↑193

[19] Johnson J., Alahi A., Fei-Fei L.. "Perceptual losses for real-time style transfer and super-resolution", *Computer Vision - ECCV 2016*, Proceedings. V. II, 14th European Conference (Amsterdam, The Netherlands, October 11–14, 2016), Lecture Notes in Computer Science, vol. **9906**, Springer, 2016, ISBN 978-3-319-46474-9, pp. 694–711. ↑199, 201

[20] Zhang J., Xiao Y., Chen G., Sun Q., Xu F., Leung C.-S.. "Histogram-guided semantic-aware colorization", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 2022 (Virtual and Singapore, May 23–27, 2022), IEEE, 2022, ISBN 978-1-6654-0541-6, pp. 2549–2553. ↑199

[21] Corley I., Lwowski J., Najafirad P.. "ZRG: A dataset for multimodal 3D residential rooftop understanding", *2024 IEEE/CVF Winter Conference on Applications of Computer Vision*, WACV 2024 (Waikoloa, HI, USA, January 03–08, 2024), IEEE, 2024, ISBN 979-8-3503-1893-7, pp. 4623–4631. doi arXiv 2304.13219 % doi ↑200

[22] Maggiori E., Tarabalka Y., Charpiat G., Alliez P.. "Can semantic labeling methods generalize to any city? The INRIA aerial image labeling benchmark", *2017 IEEE International Geoscience and Remote Sensing Symposium*, IGARSS 2017 (Fort Worth, TX, USA, July 23–28, 2017), IEEE, 2017, ISBN 978-1-5090-4951-6, pp. 3226–3229. doi ↑200

[23] Rottensteiner F., Sohn G., Jung J., Gerke M., Bailard C., Benitez S., Breitkopf U.. "The ISPRS benchmark on urban object classification and 3D building reconstruction", *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, **I**:3 (2012), pp. 293–298. doi URL ↑200

[24] Huynh-Thu Q., Ghanbari M.. "Scope of validity of PSNR in image/video quality assessment", *Electronics Letters*, **44**:13 (2008), pp. 800–801. doi ↑205

[25] Wang Z., C. Bovik A., R. Sheikh H., P. Simoncelli E.. "Image quality assessment: from error visibility to structural similarity", *IEEE Transactions on Image Processing*, **13**:4 (2004), pp. 600–612. doi ↑205

[26] Zhang R., Isola P., A. Efros A., Shechtman E., Wang O.. "The unreasonable effectiveness of deep features as a perceptual metric", *2018 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2018 (Salt Lake City, UT, USA, June 18–22, 2018), IEEE, 2018, ISBN 978-1-5386-6421-6, pp. 586–595. doi ↑205

[27] Sekrecka A., Karwowska K.. "Classical vs. machine learning-based inpainting for enhanced classification of remote sensing image", *Remote Sensing*, **17**:7 (2025), id. 1305, 36 pp. doi ↑205

[28] Heusel M., Ramsauer H., Unterthiner T., Nessler B., Hochreiter S.. "GANs trained by a two time-scale update rule converge to a local Nash equilibrium", *Advances in Neural Information Processing Systems 30*, 31st Annual Conference on Neural Information Processing Systems 2017 (Long Beach, CA, USA, December 4–9, 2017), 2017, ISBN 9781510860964, pp. 6626–6637. doi URL ↑205, 221

[29] Hasler D., E. Süsstrunk S.. "Measuring colourfulness in natural images", SPIE/IS&T Human Vision and Electronic Imaging (Santa Clara, CA, United States, 20 January 2003), Proceedings of SPIE, vol. **5007**, 2003, pp. 87–95. doi ↑215

[30] Liu Z., Hu H., Lin Y., Yao Z., Xie Z., Wei Y., Ning J., Cao Y., Zhang Z., Dong L., Wei F., Guo B.. "Swin Transformer V2: Scaling up capacity and resolution", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR 2022 (New Orleans, LA, USA, June 18–24, 2022), IEEE, 2022, ISBN 978-1-6654-6946-3, pp. 11999–12009. doi ↑220

[31] Cai H., Li J., Hu M., Gan C., Han S.. *EfficientViT: Multi-scale linear attention for high-resolution dense prediction*, 2024, 12 pp. arXiv 2205.14756 ↑220

Received 30.04.2026;
approved after reviewing 01.06.2026;
accepted for publication 10.06.2026;
published online 20.06.2026.

Recommended by *PhD. V. P. Fralenko*

Information about the authors:

Igor Victorovich Vinokurov

Candidate of Technical Sciences (PhD), Associate Professor at the Financial University under the Government of the Russian Federation. Research interests: information systems, information technologies, data processing technologies

0000-0001-8697-1032
e-mail: *igvvinokurov@fa.ru*

Georgy Mikhailovich Lapankov

Bachelor's degree graduate of the Financial University under the Government of the Russian Federation. Research interests: information systems, mobile application development, data analysis

0009-0007-0511-628X
e-mail: *goshmen2004@gmail.com*

Georgy Dmitrievich Umarov

Bachelor's degree graduate of the Financial University under the Government of the Russian Federation. Research interests: information technology, web development, data analysis

0009-0007-0364-8477
e-mail: *goshaumarov0609@mail.ru*

Authors contribution: *Igor V. Vinokurov* — 70% (development of experimental methodology, formation of configuration files, implementation of training and research of models, integration of results into the information systems of the PLC «Roscadastr» control and data processing system); *Georgy M. Lapankov* — 15% (Implementation of pre-training on ZRG); *Georgy D. Umarov* — 15% (Synthetic masks generation, visualization of model training results).

The authors declare no conflicts of interests.

RU EN ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ ISSN 2079-3316
научная статья ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И МАШИННОЕ ОБУЧЕНИЕ

УДК 004.932.75'1, 004.89
10.25209/2079-3316-2026-17-2-191-262

Roof-DeGAN: гибридная GAN с межмасштабным вниманием для восстановления областей крыш на аэрофотоснимках

Игорь Викторович **Винокуров**[1✉], Георгий Михайлович **Лапаньков**[2], Георгий Дмитриевич **Умаров**[3]

[1–3] Финансовый Университет при Правительстве Российской Федерации, Москва, Россия

[1✉] *igvvinokurov@fa.ru*

Аннотация. В работе предложена гибридная генеративно-состязательная модель Roof-DeGAN для восстановления повреждённых и скрытых участков изображений крыш на аэрофотоснимках. Архитектура сочетает Vision Transformer с плотными связями в генераторе и многоуровневый дискриминатор с межмасштабным вниманием. Модель объединяет преимущества GAN, элементов диффузионного моделирования и трансформерных механизмов. Эксперименты на данных ППК «Роскадастр» показали превосходство над современными методами: PSNR = 33,7 дБ, SSIM = 0,971, LPIPS = 0,048, FID = 17,8 при времени инференса 0,15 с на изображение. Разработанный подход обладает высокой практической ценностью для задач кадастрового учёта и обновления картографических материалов. (*Связанные тексты статьи на английском и на русском языках*)

Ключевые слова и фразы: генеративно-состязательные сети, Roof-DeGAN, восстановление изображений, инпейнтинг, аэрофотоснимки, дистанционное зондирование, реконструкция крыш

Для цитирования: Винокуров И. В., Лапаньков Г. М., Умаров Г. Д. *Roof-DeGAN: гибридная GAN с межмасштабным вниманием для восстановления областей крыш на аэрофотоснимках* // Программные системы: теория и приложения. 2026. **Т. 17**. № 2(71). С. 191–262. (*Англ.+русс.*) https://psta.psiras.ru/read/psta2026_2_191-262.pdf

Введение

Восстановление изображений и контуров зданий на аэрофотоснимках для создания актуальных карт местности и кадастровой оценки объектов капитального строительства является одной из основных задач, решаемых в ППК «Роскадастр». Эффективное решение этой задачи представляет значительный интерес для картографии, мониторинга городской инфраструктуры и градостроительного планирования [1, 2]. Однако на практике аэрофотоснимки часто содержат дефекты, вызванные атмосферными факторами (облачность, туман), временными объектами (строительная техника, транспорт, кроны деревьев) или техническими ограничениями съёмки (шумы сенсоров, низкое разрешение). Качество восстановления напрямую влияет на точность последующего автоматизированного анализа, включая семантическую и инстанс-сегментацию, построение трёхмерных моделей и оценку состояния объектов капитального строительства.

Традиционные методы восстановления изображений, такие как диффузия по Навье–Стоксу [3] и патч-матчинг на основе поиска подобия [4] демонстрируют ограниченную эффективность при работе со сложными структурами крыш. Эти подходы опираются преимущественно на низкоуровневые признаки (интенсивность пикселей, градиенты цветов и структур) и не учитывают семантику объектов, что приводит к размыванию границ, искажению геометрических форм (скаты, коньки, ендовы) и появлению неестественных текстурных артефактов. При реконструкции протяжённых повреждённых областей такие методы не способны генерировать принципиально новое содержание, что критично для аэрофотоснимков с крупными дефектами.

Современные методы глубокого обучения предлагают три основных направления решения задачи: генеративно-состязательные сети (GAN), диффузионные вероятностные модели и их гибридные комбинации.

Генеративно-состязательные сети остаются востребованными благодаря высокой скорости инференса и устойчивости к ограниченным объёмам обучающих данных. В работе [5] были представлены основы состязательного обучения, а условные GAN (*conditional GAN*, cGAN) для трансляции «изображение-в-изображение» стали базовым подходом к восстановлению. Однако классические архитектуры, такие как Pix2Pix [6], ESRGAN [7], демонстрируют падение качества при обработке крупных масок из-за ограниченной ёмкости дискриминатора и недостаточного штрафа за высокочастотные искажения, что особенно заметно на текстурах кровельных материалов.

Диффузионные вероятностные модели демонстрируют высокое качество генерации. Базовая архитектура DDPM [8] и её развитие для задач восстановления [9] позволяют достигать высокого реализма, однако требуют тысяч итераций на инференсе и больших объёмов данных. В 2025 году появились специализированные адаптации для дистанционного зондирования: SatDiff [10] на базе Stable Diffusion обеспечивает высококачественное восстановление спутниковых снимков, метод KAO [11] вводит адаптивную к ядру оптимизацию, превосходящую предыдущие подходы в структурных задачах, метод Image Characteristic-Guided [12] учитывает низкоранговые свойства изображений. Несмотря на превосходные метрики, эти модели сохраняют высокую вычислительную сложность и низкую скорость инференса.

Гибридные методы объединяют преимущества GAN и диффузии. Современные гибридные реализации GAN, такие как DeGAN [13], BD-VITGAN [14] и TAMGAN [15], интегрируют трансформеры, плотные связи и многоуровневое внимание, значительно повышая структурную согласованность и перцептивное качество в задачах удалённого зондирования. Другие подходы, такие как DSEPGAN [16], позволяют достичь баланса между скоростью, детерминированностью и качеством восстановления, что особенно важно для узкоспециализированных задач аэрофотосъёмки крыш.

Современное состояние области характеризуется фрагментарностью: большинство методов разрабатываются для общих датасетов, диффузионные модели требуют избыточных ресурсов, а гибридные подходы недостаточно адаптированы к специфике аэрофотоснимков крыш. Отсутствует специализированный метод, сочетающий устойчивость к малым выборкам, детерминированность инференса, высокое перцептивное качество и умеренную вычислительную сложность. Настоящая работа направлена на устранение этих недостатков и является продолжением исследований, описанных автором в работах [17, 18]. Предложен гибридный подход к восстановлению повреждённых областей крыш на аэрофотоснимках, основанный на условной генеративно-состязательной сети DeGAN с трансформерными блоками в генераторе и многоуровневым дискриминатором с вниманием между уровнями детализации. Составная функция потерь объединяет состязательную компоненту, попиксельное сравнение, перцептивные прмизнаки и диффузионное усиление, что позволяет достичь баланса между точностью геометрической реконструкции и естественностью текстур кровельных материалов.

В разделе 1 приводится формальная постановка задачи восстановления скрытых областей крыш. Раздел 2 описывает архитектуру Roof-DeGAN, включая строение генератора с трансформерными блоками (оснащёнными механизмом динамического разрежения внимания) и многоуровневого дискриминатора, а также составную функцию потерь. В разделе 3 представлены используемые датасеты (ZRG для предобучения и целевой датасет с синтетическими повреждениями), методы предобработки данных, а также детали двухэтапного обучения модели (предобучение на сегментацию и дообучение на восстановление) и перечень метрик количественной оценки (PSNR, SSIM, LPIPS, FID). В разделе 4 анализируется динамика обучения и сходимость модели. Раздел 5 представляет количественное сравнение предложенного подхода с современными методами. В разделе 6 проводится абляционное исследование для оценки вклада каждой компоненты функции потерь, а также влияния предобучения на ZRG. Раздел 7 анализирует качество восстановления в зависимости от площади повреждения. В разделе 8 обсуждаются ограничения и перспективные направления дальнейших исследований. Завершают статью раздел 9 и список литературы.

1. Постановка задачи

Задача восстановления повреждённых или скрытых областей крыш зданий на аэрофотоснимках может быть сформулирована как задача условной генерации изображения. Пусть $I_{gt} \in \mathbb{R}^{H\times W\times 3}$ – исходное неповреждённое RGB-изображение крыши здания размером $H \times W$ пикселей. Бинарная маска повреждённой области $M \in \{0,1\}^{H\times W}$ определяет пиксели, подлежащие восстановлению: $M_{ij} = 1$ соответствует повреждённому пикселю, $M_{ij} = 0$ – сохранным областям. Повреждённое изображение формируется как $I_{in} = I_{gt} \odot (1 - M)$, где $\odot$ – поэлементное умножение.

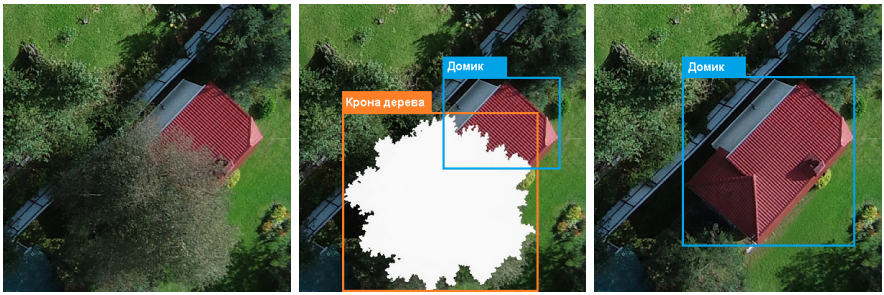
Целью является настройка параметров θ параметрической функции $G_\theta(\cdot)$ такой, что для пары (I_{in}, M) она восстанавливает полное изображение $I_{\text{rec}} = G_\theta(I_{\text{in}}, M)$, минимизируя функцию потерь генератора $\mathcal{L}_G$. Ключевой особенностью задачи является необходимость генерации нового содержания при строгом сохранении геометрической целостности архитектурных элементов крыши (скаты, коньки, ендовы) и семантической согласованности текстур кровельных материалов (черепица, металлочерепица, шифер) с окружающим контекстом, что отличает её от общих задач восстановления изображений (*inpainting*) и требует специализированных гибридных подходов (например, GAN с диффузией, учитывающей геометрические и семантические признаки).

На рисунке 1 представлена последовательность из трёх изображений, иллюстрирующая процесс восстановления скрытой области крыши – исходное повреждённое изображение, маска повреждения (крона дерева) и результат восстановления.

Необходимость решения данной задачи продиктована практическими потребностями ППК «Роскадастр». При массовой обработке аэрофотоснимков для ведения кадастрового учёта и обновления картографических материалов ключевым этапом является выделение контуров зданий и формирование цифровых карт. Однако наличие на снимках повреждённых или скрытых областей (например, закрытых кронами деревьев, тенями от облаков или временными объектами) приводит к тому, что стандартные алгоритмы сегментации и построения контуров не могут корректно определить границы крыши. Невосстановленное, зашумлённое изображение не позволяет получить достоверную бинарную маску объекта, что непосредственно сказывается на:

- точности контурных карт – на месте закрытого фрагмента крыши контур либо разрывается, либо ошибочно проводится по границе перекрывающего объекта (например, кроны дерева);
- формировании цифровых моделей рельефа и местности – искажается высотная привязка углов здания;
- кадастровом учёте – неправильно вычисляется площадь объекта, что ведёт к юридическим и экономическим последствиям.

Ручная обработка тысяч снимков с такими дефектами крайне трудоёмка и занимает значительное время. Разработка эффективной модели



(*а*) повреждённое изображение I_{in} (*б*) маска повреждения (I_{in}, M) (*в*) восстановленное изображение I_{rec}

Рисунок 1. Пример восстановления повреждённой области крыши

восстановления изображений позволяет автоматически заполнять скрытые области семантически согласованным содержанием с сохранением геометрии крыши. В результате алгоритмы картографирования получают на вход целостное изображение, что обеспечивает корректное выделение контуров, вычисление площади и формирование достоверных картографических продуктов. При этом время обработки сокращается с нескольких минут на одно изображение до долей секунды, что критически важно при работе с региональными и федеральными архивами аэрофотосъёмки.

2. Архитектура Roof-DeGAN

Предлагаемая модель представляет собой гибридную генеративно-состязательную сеть, в которой генератор восстанавливает скрытые области изображения, а дискриминатор оценивает правдоподобность полученных фрагментов. Для повышения качества восстановления применяются современные элементы: трансформерные блоки для учёта глобального контекста, плотные связи для лучшего распространения признаков и компонент диффузионного усиления для повышения стабильности и естественности текстур.

Генератор построен на основе архитектуры, сочетающей трансформерные блоки и плотные свёрточные соединения, специально адаптированной для задач восстановления спутниковых и аэрофотоснимков. Он состоит из трёх основных частей: сжимающей ветви (энкодер), промежуточного слоя (бутылочное горлышко, bottleneck) и расширяющей ветви (декодер). Между соответствующими уровнями сжимающей и расширяющей ветвей установлены skip-коннекторы, передающие низкоуровневые признаки.

Сжимающая ветвь состоит из последовательности блоков, каждый из которых включает механизм самосогласованного внимания, полносвязный слой для учёта дальних зависимостей в изображении и плотный свёрточный блок, выполняющий преобразование:

$$y = \mathrm{LeakyReLU}\left(\mathrm{BN}\left(\mathrm{Conv}_{3\times 3}(x)\right)\right) + x,$$

где

$\mathrm{Conv}_{3\times 3}(\cdot)$ – двумерная свёртка с ядром размером 3×3,
BN – пакетная нормализация,
$\mathrm{LeakyReLU}(\cdot)$ – функция активации с отрицательным наклоном.

Количество каналов признаков увеличивается от 64 до 1024 (в bottleneck), а пространственное разрешение уменьшается в два раза на каждом уровне за счёт шага свёртки равного 2.

Расширяющая ветвь симметрична сжимающей и использует транспонированные свёртки для увеличения разрешения. На каждом уровне выполняется следующее преобразование:

$$y = \text{TransposeConv}_{4\times4,\, s=2}\left(\text{Concat}\left(x_{dec}, x_{enc}\right)\right) + \text{ResidualBlock}(x),$$

где

$\text{TransposeConv}_{4\times4,\, s=2}(\cdot)$ – транспонированная свёртка с ядром 4×4 и шагом (stride) 2,

$\text{Concat}\left(x_{dec}, x_{enc}\right)$ – конкатенация (объединение по каналам) признаков текущего уровня расширяющей ветви и соответствующего уровня сжимающей ветви через skip-коннектор,

$\text{ResidualBlock}(x)$ – дополнительный блок с остаточным соединением для стабилизации обучения.

Skip-коннекторы передают детализированные низкоуровневые признаки (границы, локальные текстуры), что предотвращает размытие восстановленных областей.

Выходной слой генератора состоит из свёртки размером 1×1 с функцией активации $\tanh(\cdot)$, которая приводит значения пикселей к диапазону от -1 до 1:

$$\widehat{I} = \tanh\left(\text{Conv}_{1\times1}(z)\right),$$

где $\widehat{I}$ – восстановленное изображение, z – выход последнего блока расширяющей ветви.

Дискриминатор выполнен в виде многоуровневой сети с механизмом межмасштабного внимания. Он оценивает правдоподобность не всего изображения целиком, а одновременно на нескольких уровнях детализации (от мелких фрагментов до крупных областей), связывая информацию между разными масштабами признаков для более точной оценки согласованности восстановленных фрагментов. На вход дискриминатора подаётся конкатенация повреждённого изображения и маски повреждения. На выходе формируется набор карт вероятности разного разрешения. Каждый уровень включает свёрточную основу с функцией активации $\text{LeakyReLU}(\cdot)$ и нормализацией спектра, а также модуль межмасштабного внимания, который связывает информацию между уровнями детализации. Такой подход обеспечивает точную оценку высокочастотных деталей (текстур кровельных покрытий) и геометрических элементов крыш, а также повышает устойчивость обучения за счёт разнообразия градиентов.

Стандартные трансформерные блоки в генераторе имеют квадратичную вычислительную сложность $O(n^2)$ и не учитывают специфику аэрофотоснимков крыш, которые содержат большие однородные области (плоские скаты) наряду с мелкими деталями (черепица, стыки). Для решения этой проблемы в Roof-DeGAN предложен механизм динамического разрежения внимания (*Dynamic Sparse Attention*, DSA). На каждом уровне энкодера для каждого патча вычисляется локальная мера вариативности:

$$v_i = \mathrm{Var}(p_i) + \mathrm{Var}(\nabla p_i),$$

где p_i – значения пикселей в i-м патче, ∇p_i – градиент. Для патчей с низкой вариативностью attention вычисляется на объединении:

- предсказуемых соседних патчей (например, в окне 3×3);
- 30% случайно выбранных патчей по всему изображению.

Патчи с высокой вариативностью ($v_i \geq \tau$) обрабатываются полностью. Порог τ динамически подстраивается как перцентиль распределения $\{v_i\}$ на текущем уровне (фиксируется $\tau = P_{50}$ (50-й перцентиль, медиана распределения $\{v_i\}$)). DSA снижает вычислительную сложность с $O(n^2)$ до $O(n \cdot k)$, где $k \ll n$ – эффективный размер внимания, и позволяет обрабатывать изображения большего разрешения без потери качества на текстурированных участках.

Организация Roof-DeGAN приведена на рисунке 2.

Общая функция потерь модели $\mathcal{L}_G$ складывается из нескольких компонентов:

$$\mathcal{L}_G = \lambda_{pix}\mathcal{L}_{pix} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{diff}\mathcal{L}_{diff} + \lambda_{color}\mathcal{L}_{color}. \quad (1)$$

Пиксельная компонента ($\mathcal{L}_{pix}$) отвечает за точное совпадение яркости каждого пикселя в сохранных областях и обеспечивает геометрическую целостность изображения. Эта компонента (функция потерь $\mathcal{L}_1$) гарантирует, что модель не будет произвольно изменять яркость и форму объектов в неповреждённых частях изображения [6].

Состязательная компонента ($\mathcal{L}_{adv}$) заставляет генератор создавать настолько реалистичные фрагменты, чтобы дискриминатор не мог отличить их от настоящих. Она отвечает за реалистичность текстур и резкость границ [5].

Перцептивная компонента ($\mathcal{L}_{perc}$) сравнивает изображения не попиксельно, а в пространстве высокоуровневых признаков, извлечённых

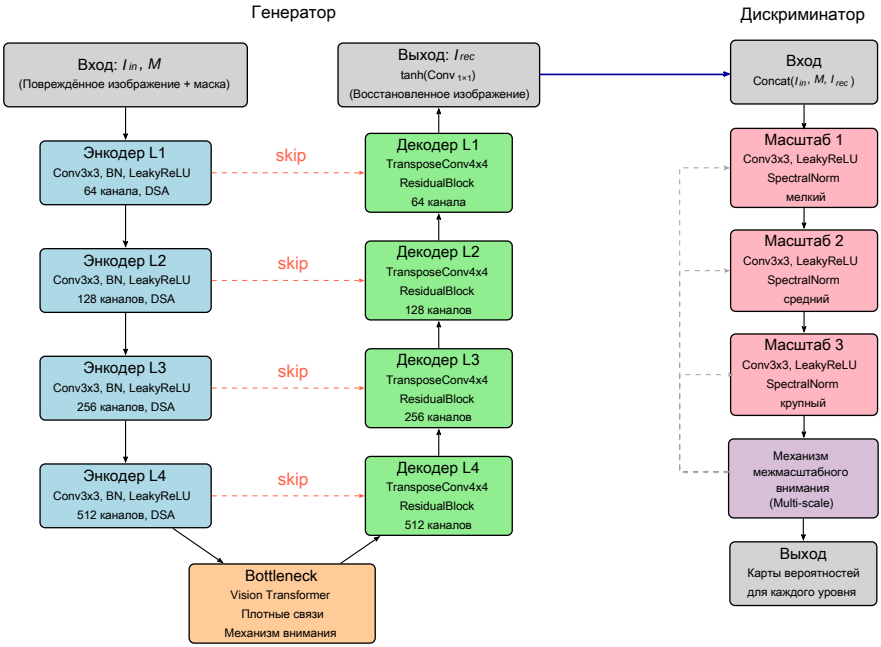


Рисунок 2. Основные компоненты модели Roof-DeGAN

предобученной свёрточной сетью, например, VGG-16 [19]. Эта компонента позволяет модели улавливать структуру объекта и текстуру материала (черепица, шифер, металл), игнорируя незначительные сдвиги пикселей, которые не влияют на визуальное восприятие человеком.

Диффузионная компонента ($\mathcal{L}_{diff}$) использует предобученную модель шумоподавления DDPM [8] в качестве эксперта по реалистичности. Она заставляет восстановленное изображение вести себя так же, как эталон, в процессе диффузионного зашумления и восстановления. Подход основан на идеях, изложенных в работах [9, 12], и позволяет эффективно подавлять характерные для GAN высокочастотные артефакты, а также стабилизировать обучение.

Компонента цветовой согласованности ($\mathcal{L}_{color}$) штрафует модель за появление неестественных оттенков, сравнивая распределения цветов в восстановленном и эталонном изображениях. Реализация основана на дифференцируемой гистограммной потере [20]. Эта компонента сохраняет общую цветовую гамму, свойственную конкретному типу кровельного

покрытия (например, терракотовый цвет черепицы или серый оттенок шифера), предотвращая появление «кислотных» или выцветших тонов [12].

Лучшие значения коэффициентов λ были подобраны экспериментально в ходе исследования функции потерь (см. раздел 6).

3. Детали обучения и оценка точности

3.1. Предобучение на ZRG (этап сегментации)

Для формирования начальных представлений о геометрии и текстурах кровельных покрытий использовался приватный датасет Zeitview Rooftop Geometry (ZRG) [21].[1] Датасет содержит 22 334 аннотированных RGB-изображения жилых домов, полученных с беспилотных летательных аппаратов на высоте 3–5 метров над крышами. Разрешение снимков составляет <1 см/пиксель, что позволяет различать мелкие элементы кровли (отдельные черепицы, стыки, вентиляционные выходы). Каждое изображение сопровождается:

- сегментационной маской крыши (бинарная маска «крыша/фон»);
- 3D-каркасом (полигональная модель граней крыши);
- информацией о наличии нависающей растительности и теней (естественные окклюзии).

Выбор ZRG обусловлен следующими преимуществами по сравнению с классическими датасетами аэрофотоснимков (Inria [22], ISPRS Potsdam [23]). В отличие от Inria и ISPRS Potsdam, которые содержат только бинарную разметку «здание/не здание», ZRG предоставляет детальную сегментационную маску крыши и 3D-каркас. Разрешение ZRG (<1 см/пиксель) значительно превосходит разрешение Inria (30 см/пиксель) и ISPRS Potsdam (5 см/пиксель), что позволяет различать мелкие элементы кровли. Кроме того, ZRG включает естественные окклюзии (тени, нависающие деревья), отсутствующие в классических датасетах, ориентирован на жилую и пригородную застройку (в отличие от городской в Inria) и насчитывает более 22 000 аннотированных объектов против 38 патчей в ISPRS Potsdam.

[1] ППК «Роскадастр» использует датасет на законных основаниях

Генератор Roof-DeGAN предварительно обучался на датасете ZRG в режиме бинарной сегментации крыш. Архитектура на этом этапе включала энкодер-декодер с трансформерными блоками и плотными skip-соединениями (как в целевой модели), но без состязательной компоненты. Функция потерь на этапе предобучения:

$$L_{pretrain} = \lambda_{pix} L_{pix} + \lambda_{perc} L_{perc},$$

где

L_{pix} – $\mathcal{L}_1$-потеря (попиксельное сравнение с маской крыши),
L_{perc} – перцептивная потеря на признаках VGG-16 [19].

Коэффициенты: $\lambda_{pix} = 1{,}0$, $\lambda_{perc} = 0{,}05$.

Обучение продолжалось в течение 65 эпох с использованием оптимизатора Adam ($lr = 10^{-4}$, $\beta_1 = 0{,}5$, $\beta_2 = 0{,}999$), размер мини-батча – 16 изображений. Веса дискриминатора на этом этапе не инициализировались.

3.2. Перенос весов и дообучение (этап восстановления)

После завершения предобучения веса энкодера и декодера копировались в генератор целевой модели Roof-DeGAN. Выходной слой генератора заменялся: вместо одного канала (бинарная маска) устанавливалось три канала (RGB), его веса инициализировались случайно. Дискриминатор инициализировался случайным образом.

Дообучение проводилось на целевом датасете (1600 изображений с синтетическими масками повреждений) с использованием полной функции потерь (1). Оптимальные коэффициенты: $\lambda_{pix} = 1{,}0$, $\lambda_{adv} = 0{,}1$, $\lambda_{perc} = 0{,}05$, $\lambda_{diff} = 0{,}01$, $\lambda_{color} = 0{,}05$. Обучение продолжалось в течение 30 эпох с ранней остановкой (обучение прекращалось, если функция потерь на валидационной выборке (200 изображений) не уменьшалась в течение 5 последовательных эпох). Согласно динамике обучения (рисунок 4 и 6), наилучшие метрики достигнуты на 25-й эпохе, после чего на 26-й эпохе зафиксировано резкое ухудшение (коллапс состязательного обучения). Поэтому в качестве итоговой модели сохранены веса, соответствующие 25-й эпохе.

Размер мини-батча составил 32 изображения. Благодаря использованию смешанной точности (FP16) и механизма Flash Attention, обучение стабильно выполнялось на GPU NVIDIA A100 80 ГБ. Выбор батча 32 обеспечивает баланс между вычислительной эффективностью и гладкостью градиентов, необходимой для состязательного обучения.

3.3. Целевой датасет и генерация синтетических масок повреждений

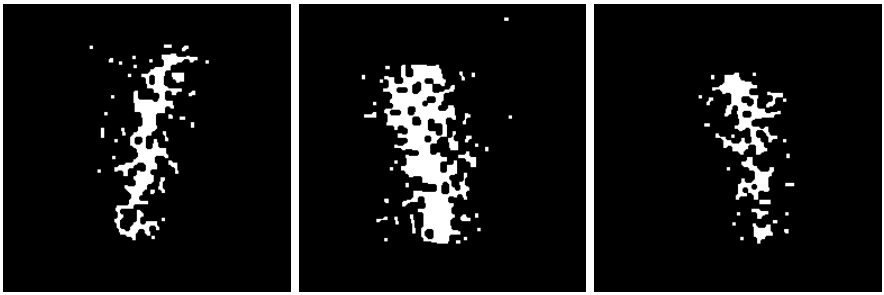
Целевой датасет сформирован на основе архивов аэрофотосъёмки ППК «Роскадастр»[2] и включает 2000 RGB-изображений крыш зданий (1600 – для обучения, 200 – для валидации, 200 – для тестирования). Все изображения получены с беспилотных летательных аппаратов (БПЛА) на высоте 30–50 м над уровнем крыш, что для типовых садово-дачных построек (≈8×8 м) обеспечивает пространственное разрешение ≈3 см/пиксель. Размер каждого изображения после предобработки составляет 256×256 пикселей, что соответствует одному зданию целиком с небольшим окружающим фоном. Датасет охватывает следующие типы кровельных материалов: металлочерепица (около 35%), керамическая черепица (25%), шифер (20%), битумная черепица (10%) и прочие покрытия (рулонные покрытия, плоские крыши, медь) – 10%. По типу застройки снимки распределены следующим образом: городская плотная (10%), пригородная (15%) и садово-дачная (75%). Все изображения проходят ручную верификацию на отсутствие глобальных дефектов (разрывов более 50% площади, сильных атмосферных искажений, например дымки или бликов), что обеспечивает чистоту эталонных данных.

Поскольку реальные аэрофотоснимки с точной маской «скрытая область» (например, под кроной дерева или строительной техникой) получать трудоёмко и дорого, в работе используется синтетическая генерация масок повреждений. В задачах восстановления изображений такой подход является общепринятым и позволяет контролировать форму, размер и положение дефектов. Маска повреждения $M \in \{0,1\}^{H \times W}$ формируется для каждого обучающего примера по следующему алгоритму. Сначала с заданными вероятностями (40%, 30% и 30% соответственно) выбирается один из трёх типов повреждения: прямоугольное/многоугольное (имитация закрытия строительной техникой или временными сооружениями), реалистичная крона дерева или нерегулярное составное (моделирование теней облаков или спутанных объектов).

Особое внимание уделено генерации масок крон деревьев, которые в отличие от примитивных геометрических фигур (эллипсов или прямоугольников) воспроизводят характерные особенности реальной древесной растительности. В разработанном методе предусмотрены следующие типы крон (примеры приведены на рисунке 3):

[2] Собственность ППК «Роскадастр»

(а) Ветвистая крона



(б) Пирамидальная крона

(в) Раскидистая крона

Рисунок 3. Примеры синтетических масок крон деревье

- Ветвистые (старый дуб, вяз) – кроны с выраженными крупными ветвями, выступающими за пределы основной массы;
- Пирамидальные (ель) – классические конусовидные кроны с характерной формой и «пушистой» текстурой;
- Раскидистые (сосна) – неправильные, асимметричные кроны с неравномерным ростом ветвей.

Для каждого типа кроны задаются следующие параметры: площадь A_m выбирается из равномерного распределения в одном из трёх диапазонов – 10–15%, 20–25% или 30–35% от общей площади изображения, что позволяет оценить устойчивость модели к различной степени перекрытия (см. раздел 7). Для ветвистых форм дополнительно регулируется количество и длина ветвей; для пирамидальных – угол наклона ствола и степень «пушистости»; для раскидистых – степень асимметрии и количество центров роста кроны. Местоположение маски с равной вероятностью выбирается в центре, на краю или в углу крыши, чтобы модель обучалась восстанавливать как изолированные дефекты, так и краевые разрывы.

После генерации формы маска подвергается постобработке: сначала применяется морфологическое открытие с ядром 3×3 для удаления изолированных ложных пикселей на границах маски, затем – гауссово размытие с параметром $\sigma = 1$ и последующая бинаризация по порогу $0,5$. Это создаёт плавные, но чёткие границы перехода между повреждённой и сохранной областями, предотвращая появление резких ступенек яркости, которые модель могла бы использовать как «простой ключ» для обхода обучения. Дополнительно для масок крон деревьев с вероятностью 50% генерируются внутренние просветы (имитация промежутков между ветвями), а с вероятностью 40% – торчащие ветки, что повышает реалистичность и усложняет задачу восстановления.

Повреждённое изображение формируется как $I_{in} = I_{gt} \odot (1 - M)$, то есть пиксели, соответствующие маске, обнуляются, а сохранные области остаются неизменными. Генерация масок выполняется на лету (*on-the-fly*) в процессе загрузки данных (размер батча составляет 32 изображения). Для валидационной и тестовой выборок маски и повреждённые изображения фиксируются и сохраняются один раз, что обеспечивает воспроизводимость сравнения между разными моделями и конфигурациями.

Предложенный подход к генерации масок отражает наиболее частые на практике случаи перекрытия крыш: древесная растительность (реалистичные кроны деревьев различных типов площадью 15–35%), техника и временные объекты (прямоугольные маски 10–20%), облака и тени (крупные нерегулярные маски 20–35%). К ограничениям метода следует отнести отсутствие имитации полутоновых и градиентных перекрытий (например, ажурная тень от листвы), а также отсутствие масок с разрывом более 50% площади (такие случаи требуют уже не восстановления, а генерации «с нуля» и выходят за рамки данной задачи). Тем не менее, как показано в разделе 7 (рисунок 9), уже при 35% перекрытия качество восстановления может становиться неприемлемым, что аргументирует выбор указанных диапазонов площадей повреждений.

3.4. Метрики оценки качества

Для количественной оценки качества восстановления использовались следующие метрики: пиковое отношение сигнала к шуму (PSNR) [24], индекс структурного сходства (SSIM) [25] и метрика перцептивного сходства (LPIPS) [26] на основе признаков предобученной сети VGG-16. Все метрики рассчитывались исключительно по пикселям, заданным маской повреждения M, что обеспечивает корректное сравнение методов в задачах восстановления. Выбор LPIPS обусловлен высокой корреляцией глубоких признаков VGG-16 с человеческим восприятием текстур и форм кровельных материалов [27].

Дополнительно для оценки качества генерации текстур применялась метрика дистанции Фреше между распределениями inception-векторов (FID) [28], вычисляемая между распределениями восстановленных и эталонных изображений на тестовой выборке. Низкое значение FID свидетельствует о том, что сгенерированные фрагменты не только похожи на оригиналы попиксельно, но и статистически неотличимы от реальных изображений крыш.

Для оценки точности восстановления геометрических контуров крыш (скаты, коньки, ендовы) используется метрика Boundary F1 [1]. Границы выделяются детектором Канни (пороги 50, 150), после чего вычисляется F1-мера с допуском 3 пикселя. Метрика рассчитывается только в области повреждения M и характеризует пригодность восстановленного изображения для автоматического контурирования в кадастровых системах.

Для оценки надёжности метрик качества использован метод бутстрэпа (1000 итераций ресэмплинга с возвращением на тестовой выборке из 200 изображений). Результаты представлены как среднее $\pm$ стандартное отклонение; для метрики FID дополнительно приведены 95% доверительные интервалы.

Для визуальной валидации результатов использовался также метод экспертных оценок: три специалиста в области кадастрового учёта проводили сравнительный анализ восстановленных изображений по шкале от 1 до 5, оценивая геометрическую точность и реалистичность текстур. Усреднённая экспертная оценка (*Mean Opinion Score*, MOS) позволила подтвердить корреляцию объективных метрик с субъективным восприятием качества. Все метрики вычислялись как на всей тестовой выборке (200 изображений), так и отдельно для трёх групп повреждений (10–15%, 20–25%, 30–35%), что позволило оценить эффективность работы модели для различных степеней перекрытия крыш.

Эксперименты проводились на вычислительном кластере со следующей конфигурацией: GPU NVIDIA Tesla A100 SXM (80 ГБ памяти), CPU Intel Xeon Gold 6248R (24 ядра), оперативная память 128 ГБ DDR4. Используемое программное обеспечение: Python 3.12, PyTorch 2.5 с поддержкой CUDA 12.x. Для воспроизведения базовой функциональности и экспериментов с небольшими датасетами (предобучение на Inria) также подготовлена упрощённая версия модели в виде интерактивного блокнота Jupyter в облачной среде Google Colab.

4. Динамика обучения и сходимость модели

На рисунке 4 представлены кривые изменения основных компонентов функции потерь на валидационной выборке в процессе обучения модели.

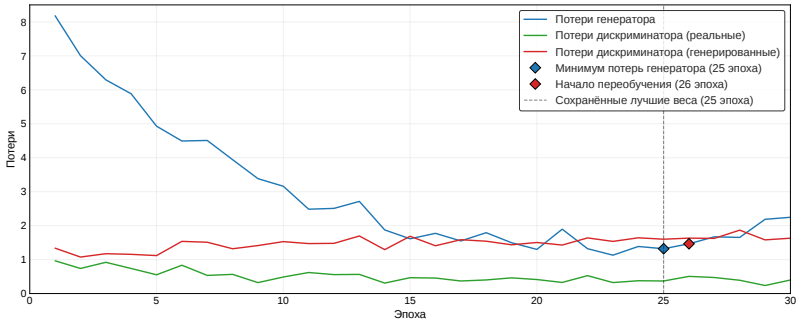


Рисунок 4. Динамика изменения функций потерь генератора и дискриминатора на валидационной выборке при обучении Roof-DeGAN

В первые 10 эпох наблюдается быстрое снижение потерь генератора (с 7,82 до 3,94), что соответствует начальной фазе обучения, когда модель осваивает грубую структуру изображений и общие геометрические особенности крыш. Потери дискриминатора на реальных изображениях монотонно уменьшаются (с 0,89 до 0,58), что указывает на улучшение способности различать эталонные образы. Одновременно потери дискриминатора на сгенерированных изображениях возрастают (с 1,12 до 1,45), отражая типичное для состязательного обучения усиление градиентов против генератора.

В интервале 10–20 эпох темп снижения потерь генератора замедляется, модель начинает уточнять текстуры кровельных материалов и границы архитектурных элементов. Потери дискриминатора на реальных изображениях продолжают плавно снижаться (до 0,41 к 20-й эпохе), а на

сгенерированных – растут (до 1,62), что свидетельствует о сохранении баланса между генератором и дискриминатором.

В интервале 20–25 эпох наблюдается дальнейшее улучшение качества: потери генератора снижаются до минимального значения 1,32 к 25-й эпохе, потери дискриминатора на реальных изображениях стабилизируются на уровне 0,37, на сгенерированных – 1,60. Метрики качества достигают пиковых значений: PSNR – 34,5 дБ, SSIM – 0,972, LPIPS – 0,052, FID – 17,6.

На 26-й эпохе происходит резкое ухудшение: потери генератора возрастают до 2,38, а метрики качества демонстрируют спад (PSNR снижается до 33,42 дБ, LPIPS возрастает до 0,058, FID – до 18,7). Дальнейшее обучение до 30 эпох не приводит к восстановлению качества: потери генератора колеблются в диапазоне 2,35–2,41, метрики остаются на худших значениях. Это послужило основанием для применения ранней остановки (*early stopping*) на 26-й эпохе с сохранением лучшей модели, полученной на 25-й эпохе. В работе не применялись распространённые техники стабилизации GAN (R1-регуляризация и спектральная нормализация). Стабильность достигалась преимущественно архитектурными решениями (DSA, межмасштабный дискриминатор) и композитной функцией потерь. Применение указанных методов – перспективное направление дальнейшей работы.

На рисунке 6 показана динамика изменения метрик качества PSNR, SSIM, LPIPS и FID на валидационной выборке в процессе обучения модели. Анализ этого рисунка подтверждает выводы, полученные на основе кривых потерь. Метрики PSNR и SSIM активно растут до 20-й эпохи, после чего темп роста замедляется. Максимальные значения достигаются на 25-й эпохе: PSNR = 34,5 дБ, SSIM = 0,972. Метрика LPIPS снижается до 0,052, а FID – до 17,6, что указывает на значительное улучшение перцептивного качества и естественности текстур кровельных покрытий. На 26-й эпохе фиксируется ухудшение всех метрик: PSNR снижается до 33,42 дБ, SSIM – до 0,970, LPIPS возрастает до 0,058, FID – до 18,7.

В интервале 27–30 эпох наблюдается плато на ухудшенных значениях, что свидетельствует о начале переобучения – модель начинает чрезмерно подстраиваться под текстуры обучающей выборки в ущерб обобщающей способности на новых изображениях.

Полученные результаты подтверждают эффективность выбранных техник стабилизации (ранняя остановка, сглаживание меток) и демонстрируют, что предложенный гибридный подход Roof-DeGAN достигает баланса между скоростью сходимости, перцептивным качеством и устойчивостью к ограниченному объёму обучающих данных.

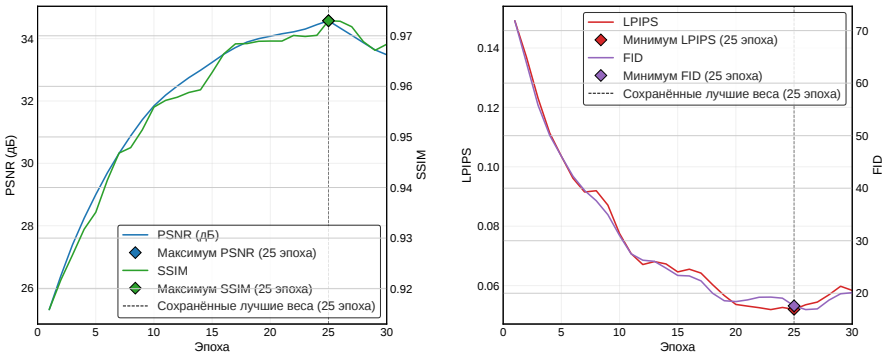


Рисунок 5. Динамика изменения геометрических (PSNR, SSIM) и перцептивных (LPIPS, FID) метрик на валидационной выборке в процессе обучения Roof-DeGAN

5. Количественное сравнение с современными методами

Для объективной оценки эффективности предложенного подхода проведено сравнение с рядом современных методов восстановления изображений, включая традиционные алгоритмы, чистые диффузионные модели и гибридные GAN-архитектуры. Тестирование выполнялось на идентичной тестовой выборке из 200 изображений с одинаковыми синтетическими и естественными масками повреждений площадью 15–25%. Все методы исследовались в стандартных конфигурациях на одном оборудовании (GPU NVIDIA A100 SXM и CPU Intel Xeon Gold 6248R). Результаты представлены в таблице 1. Символ ↑ в заголовке таблицы обозначает, что большее значение лучше, символ ↓ – меньшее значение лучше. Время инференса указано для изображений размером 256×256 пикселей.

Анализ этой таблицы позволяет сделать следующие выводы.

Превосходство над традиционными методами. Roof-DeGAN значительно опережает классические алгоритмы Navier-Stokes и PatchMatch по всем метрикам: выигрыш в PSNR составляет 8,0–9,7 дБ, а FID снижается в 8–10 раз. Это визуально соответствует переходу от артефактных и размытых реконструкций к детализированным изображениям с сохранением геометрии и текстур кровельных покрытий.

Таблица 1. Сравнение Roof-DeGAN с современными методами на тестовой выборке ($N = 200$)

Методы	PSNR↑	SSIM↑	LPIPS↓	FID↓	**П**[1]	t[2]
Традиционные						
Navier–Stokes [3]	24,18	0,742	0,312	187,3	–	0,8
PatchMatch [4]	25,92	0,801	0,245	142,8	–	2,3
Диффузионные						
DDPM [8]	28,45	0,873	0,118	78,6	112	28,5
SatDiff [10]	30,45	0,941	0,065	38,2	189	41,3
KAO [11]	31,12	0,952	0,058	32,7	205	38,9
GAN-методы						
Pix2Pix [6]	28,92	0,885	0,142	92,5	54	**0,12**
ESRGAN [7]	29,34	0,902	0,098	67,3	67	0,18
DeGAN [13]	31,85	0,958	0,082	29,4	62	0,14
Новый метод						
Roof-DeGAN	$\mathbf{33,7} \pm 1,2$	$\mathbf{0,971} \pm 0,008$	$\mathbf{0,048} \pm 0,011$	$\mathbf{17,8} \pm 2,4$	**48**	0,15

[1] П – количество параметров модели в миллионах;
[2] t – время инференса на одно изображение в секундах.

Сравнение с диффузионными моделями. Чистые диффузионные модели (DDPM, SatDiff, KAO) демонстрируют высокое качество, однако требуют десятков секунд на инференс и значительного объёма обучающих данных. Предложенный гибридный метод превосходит их по PSNR на 2,8–5,5 дБ и достигает существенно лучшего FID (17,8 против 32,7–78,6) при времени инференса в 250–270 раз меньшем (0,15 с. против 38–41 с.). Это делает подход практически применимым для обработки больших массивов аэрофотоснимков в реальном времени.

Сравнение с GAN-архитектурами. По сравнению с базовыми GAN-моделями (Pix2Pix, ESRGAN) и другими гибридными подходами, такими как DeGAN, предложенный метод показывает прирост PSNR до 5,0 дБ и значительное улучшение FID (в 1,7–5,2 раза). Значение LPIPS = 0,048 подтверждает высокое перцептивное качество восстановленных текстур, превосходящее показатели сравниваемых GAN-архитектур (в нашем случае это Pix2Pix = 0,142; ESRGAN = 0,098; DeGAN = 0,082) – улучшение на 41–66% относительно ближайшего конкурента. При этом модель имеет меньшее количество параметров (48 M против 54–205 M), что обеспечивает экономию вычислительных ресурсов. Превосходство Roof-DeGAN над ближайшими конкурентами является статистически значимым: 95% доверительные интервалы метрик не перекрываются. Например, разница в PSNR между Roof-DeGAN (33,7 ± 1,2 дБ) и KAO (31,1 ± 1,8 дБ) составляет 2,6 дБ при непересекающихся интервалах, что подтверждает устойчивость преимущества предложенного метода.

Полученные результаты демонстрируют, что интеграция трансформерных блоков, многоуровневого внимания и диффузионного усиления позволяет предложенному методу достигать наилучшего баланса между качеством восстановления и вычислительной эффективностью среди рассмотренных подходов.

6. Исследование влияния компонент функции потерь

Для количественной оценки вклада каждой компоненты композитной функции потерь проведён сравнительный анализ пяти конфигураций модели, различающихся набором оптимизируемых критериев. Пять конфигураций модели обучались с различными комбинациями компонентов до достижения оптимальной эпохи по критерию ранней остановки. Результаты представлены в таблице 2, лучшие показатели выделены красным.

Таблица 2. Абляционное исследование основных компонентов функции потерь на валидационной выборке

Конфигурация (компоненты из (1))	PSNR↑	SSIM↑	LPIPS↓	FID↓	**Визуальные особенности**
Без предобучения (все компоненты из (1))	28,34	0,912	0,112	54,2	Заметные артефакты, текстуры сглажены
B ($\lambda_{pix}\mathcal{L}_{pix}$)	29,84	0,931	0,098	87,4	Размытые текстуры, сглаженные границы
A ($\lambda_{pix}\mathcal{L}_{pix} + \lambda_{adv}\mathcal{L}_{adv}$)	31,12	0,952	0,082	48,2	Чёткие границы, локальные текстурные артефакты
P ($\lambda_{pix}\mathcal{L}_{pix} + \lambda_{perc}\mathcal{L}_{perc}$)	31,45	0,958	0,068	39,7	Естественные текстуры, избыточное сглаживание
D ($\lambda_{pix}\mathcal{L}_{pix} + \lambda_{adv}\mathcal{L}_{adv} + + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{diff}\mathcal{L}_{diff}$)	32,18	0,965	0,058	28,9	Хорошая геометрия и текстура, небольшие цветовые сдвиги
C ($\lambda_{pix}\mathcal{L}_{pix} + \lambda_{adv}\mathcal{L}_{adv} + + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{diff}\mathcal{L}_{diff} + + \lambda_{color}\mathcal{L}_{color}$)	34,5	0,972	0,052	17,6	Наиболее реалистичные текстуры, геометрия и цветовая согласованность

Анализ таблицы 2 показывает следующее. Базовая попиксельная конфигурация B (Baseline) обеспечивает геометрическую целостность (положение скатов, коньков, ендов), но приводит к характерной размытости текстур кровельных покрытий (PSNR = 29,84 дБ, LPIPS = 0,098).

Добавление состязательной компоненты A (Adversarial) существенно повышает чёткость границ и добавляет высокочастотные детали, однако без перцептивной и диффузионной составляющих возникают локальные артефакты (LPIPS = 0,082, FID остаётся высоким).

Включение перцептивной компоненты P (Perceptual) без состязательной части позволяет достичь значения LPIPS = 0,068, однако визуально изображения выглядят излишне сглаженными из-за отсутствия высокочастотных деталей, которые вносит состязательная компонента.

Наилучшее значение LPIPS (0,058) среди неполных конфигураций достигается при совместном использовании состязательной, перцептивной и диффузионной компонент D (Diffusion). Добавление диффузионного усиления значительно улучшает структурную согласованность и снижает FID (до 28,9), однако наблюдаются небольшие цветовые сдвиги.

Полная конфигурация C (Color), включающая все пять компонентов, объединяет их преимущества и достигает на валидационной выборке лучших значений геометрических и перцептивных метрик точности PSNR = 34,5 дБ, SSIM = 0,972, LPIPS = 0,052, FID = 17,6.

Сравнение конфигурации «Без предобучения» и полной конфигурации C показывает, что предложенный двухэтапный подход (предобучение на ZRG с последующим дообучением) обеспечивает прирост PSNR на 6,16 дБ, снижение LPIPS на 0,060 и снижение FID на 36,6 по сравнению с обучением с нуля. Это подтверждает эффективность переноса весов из задачи сегментации крыш в задачу восстановления изображений.

Абляционное исследование (таблица 2) подтвердило необходимость использования всех пяти компонент функции потерь. Для достижения наилучшего баланса между геометрической точностью и перцептивным качеством была проведена настройка весовых коэффициентов λ. Оптимальные значения, полученные в ходе гиперпараметрической оптимизации на валидационной выборке, составили:

- $\lambda_{pix} = 1{,}0$ (базовое пиксельное соответствие),
- $\lambda_{adv} = 0{,}1$ (реалистичность),
- $\lambda_{perc} = 0{,}05$ (коррекция текстур),
- $\lambda_{diff} = 0{,}01$ (сохранение геометрии),

• $\lambda_{color} = 0{,}05$ (цветовой баланс).

На рисунке 6 представлена динамика изменения метрик точности на валидационной выборке.

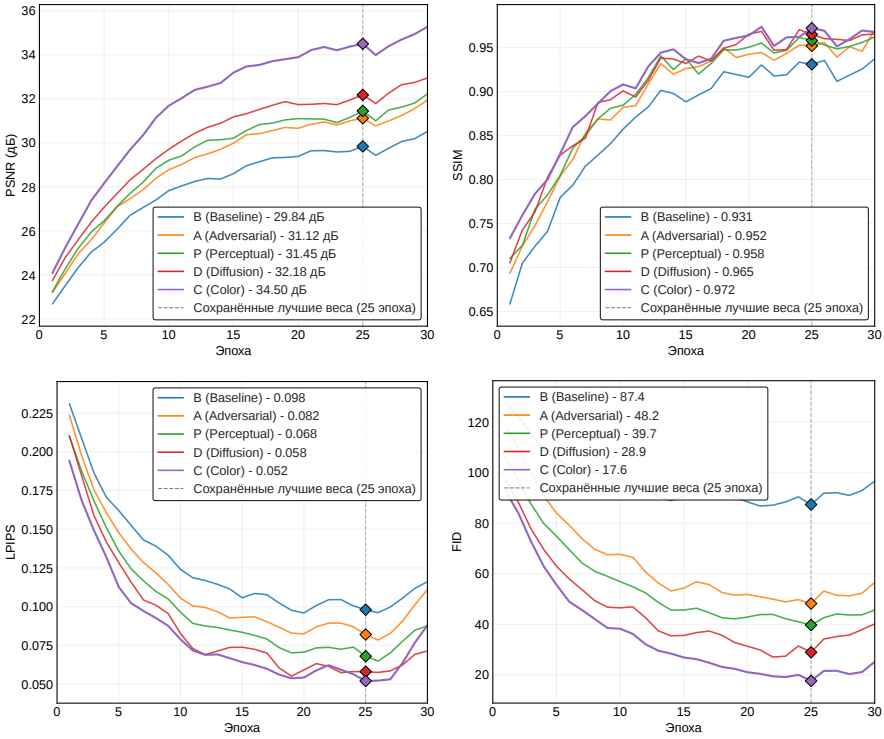


Рисунок 6. Динамика изменения геометрических (PSNR,SSIM) и перцептивных (LPIPS,FID) метрик на валидационной выборке

Пиксельная компонента ($\lambda_{pix} = 1{,}0$). Данная компонента выполняет роль базового регуляризатора, обеспечивающего сохранение неизменных областей изображения и грубое соответствие восстановленных фрагментов эталону по яркости. Значение $\lambda_{pix} = 1{,}0$ выбрано в качестве референсного, поскольку $\mathcal{L}_{pix}$ ($\mathcal{L}_1$-потеря) имеет естественный масштаб, сопоставимый с суммарным вкладом остальных компонент. Уменьшение λ_{pix} до 0,5 приводит к заметному размытию границ архитектурных элементов (PSNR снижается на 1,2 дБ), а увеличение до 2,0 подавляет состязательную и

перцептивную компоненты, делая текстуры излишне сглаженными.

Состязательная компонента ($\lambda_{adv} = 0{,}1$). GAN-обучение склонно к нестабильности при высоких значениях состязательной компоненты. Экспериментально установлено, что при $\lambda_{adv} > 0{,}2$ дискриминатор слишком быстро сходится, порождая ослабевающие градиенты для генератора и приводя к коллапсу мод уже после 10–12 эпох. При $\lambda_{adv} < 0{,}05$ влияние состязательной составляющей становится незначительным: модель генерирует геометрически корректные, но излишне гладкие текстуры (LPIPS $> 0{,}07$). Значение $\lambda_{adv} = 0{,}1$ обеспечивает устойчивое состязательное равновесие, при котором дискриминатор остаётся достаточно «сильным», чтобы предоставлять информативный градиент, но не подавляет генератор. Следует отметить, что увеличение λ_{adv} с $0,05$ до $0,1$ повышает Boundary F1 с $0,89$ до $0,93$, подтверждая важность состязательной компоненты для чёткости геометрических границ.

Перцептивная компонента ($\lambda_{perc} = 0{,}05$). Перцептивная потеря, вычисляемая на признаках VGG-16, имеет существенно больший масштаб по сравнению с $\mathcal{L}_1$-потерей. Прямое использование $\lambda_{\text{perc}} = 1{,}0$ приводит к доминированию этой компоненты и появлению характерных артефактов — чрезмерной текстурированности и «галлюцинаций» мелких деталей, отсутствующих в эталоне. Понижение коэффициента до 0,05 позволяет сохранить положительный эффект перцептивного обучения (естественность текстур черепицы и шифера) без перекоса общей функции потерь. Абляционное исследование (таблица 2) подтверждает, что исключение $\mathcal{L}_{perc}$ увеличивает LPIPS на 0,030, а удвоение λ_{perc} до 0,1 не даёт значимого улучшения, но замедляет сходимость.

Диффузионная компонента ($\lambda_{diff} = 0{,}01$). Диффузионное усиление основано на предобученной модели DDPM, выступающей в роли «эксперта по реалистичности». Масштаб диффузионной потери существенно варьируется в зависимости от уровня шума и текущего состояния генератора. Значение $\lambda_{diff} = 0{,}01$ подобрано таким образом, чтобы компонента оказывала стабилизирующее влияние (подавление высокочастотных GAN-артефактов, снижение FID) без доминирования над пиксельной и состязательной составляющими. При $\lambda_{diff} > 0{,}05$ модель начинает копировать текстурные особенности предобученной диффузионной модели, что приводит к избыточному сглаживанию мелких деталей кровли. При $\lambda_{diff} < 0{,}005$ положительный эффект диффузионного усиления становится статистически незначимым.

Компонента цветовой согласованности ($\lambda_{color} = 0{,}05$). Цветовая потеря, основанная на дифференцируемой гистограмме, имеет масштаб, чувствительный к размеру изображения и количеству бинов гистограммы. Для изображений 256×256 пикселей коэффициент 0,05 является оптимальным: он эффективно подавляет неестественные цветовые сдвиги (например, появление «кислотных» оттенков при восстановлении терракотовой черепицы), но не приводит к усреднению цветовых кластеров. При $\lambda_{color} = 0{,}1$ наблюдается лёгкое «выцветание» насыщенных цветов кровельных материалов (снижение цветового разнообразия по метрике Colorfulness Index [**29**] на 12%).

Итоговый баланс. Таким образом, выбранные коэффициенты обеспечивают сбалансированный вклад каждой компоненты в общую функцию потерь, что подтверждается результатами абляционного исследования: полная конфигурация достигает на валидационной выборке PSNR = 34,5 дБ, SSIM = 0,972, LPIPS = 0,052 и FID = 17,6, превосходя все неполные конфигурации (таблица 2) по совокупности геометрических и перцептивных метрик.

7. Анализ качества восстановления в зависимости от площади повреждения

Для оценки устойчивости модели к различным масштабам повреждений тестовая выборка из 200 изображений была разделена на три группы: малые повреждения (10–15% площади), средние (20–25%) и обширные (30–35%). Малые повреждения составляют порядка 85% от общего количества изображений в тестовой выборке, средние и обширные – 10% и 5% соответственно.

Результаты сравнения с современными методами приведены в таблице 3. Лучшие показатели выделены красным. Как видно из таблицы, Roof-DeGAN стабильно превосходит все сравниваемые подходы во всех диапазонах повреждений. При малых повреждениях (10–15%) выигрыш в PSNR составляет 1,5–9,7 дБ по сравнению с традиционными методами и 1,5–2,1 дБ относительно современных диффузионных моделей (SatDiff, KAO). При этом SSIM улучшается до 0,986 против 0,742–0,958 у конкурентов, а FID снижается до 18,2 против 32,7–187,3.

При средних повреждениях (20–25%) преимущество усиливается: прирост PSNR достигает 2,0–8,3 дБ, SSIM растёт до 0,971 против 0,718–0,952, LPIPS снижается до 0,059 против 0,058–0,328 у конкурентов, а FID уменьшается до 22,5 против 35,8–192,6.

Таблица 3. Качество восстановления в зависимости от площади повреждения

Методы	10–15%				20–25%				30–35%			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Традиционные методы												
Navier-Stokes [3]	24,18	0,742	0,312	187,3	23,45	0,718	0,328	192,6	22,10	0,682	0,356	205,4
PatchMatch [4]	25,92	0,801	0,245	142,8	25,12	0,778	0,262	148,5	23,85	0,741	0,289	162,1
Диффузионные модели												
DDPM [8]	28,45	0,873	0,118	78,6	27,80	0,852	0,132	82,4	26,15	0,814	0,158	89,7
SatDiff [10]	31,78	0,949	0,068	38,2	30,45	0,941	0,065	41,5	28,92	0,912	0,092	52,3
KAO [11]	32,41	0,958	0,061	32,7	31,12	0,952	0,058	35,8	29,67	0,923	0,085	44,1
Гибридные и GAN-методы												
Pix2Pix [6]	28,92	0,885	0,142	92,5	28,15	0,865	0,156	98,2	26,78	0,828	0,182	110,6
ESRGAN [7]	29,34	0,902	0,098	67,3	28,65	0,882	0,112	72,1	27,40	0,845	0,138	84,9
DeGAN базовый [13]	31,85	0,958	0,082	29,4	31,02	0,948	0,088	33,7	29,78	0,922	0,102	41,2
Новый метод												
Roof-DeGAN	33,87	0,986	0,045	18,2	33,42	0,971	0,059	22,5	31,25	0,948	0,074	29,8

Особенно показательно поведение при обширных повреждениях (30–35%): предложенный метод сохраняет высокое качество (PSNR = 31,25 дБ, SSIM = 0,948, LPIPS = 0,074, FID = 29,8), тогда как традиционные методы падают до 22,10–23,85 дБ (FID до 205,4), а чистые диффузионные модели — до 28,92–29,67 дБ (FID до 44,1–89,7). Выигрыш в PSNR по сравнению с лучшими конкурентами (KAO, DeGAN) составляет 1,6–2,3 дБ, SSIM улучшается на 0,025–0,026, LPIPS сопоставим (0,074 против 0,058–0,085 у конкурентов), а FID уменьшается на 11,3–14,4 единицы. Результаты сравнения геометрических и перцептивных метрик модели Roof-DeGAN с лучшими конкурентами приведены на рисунке 7.

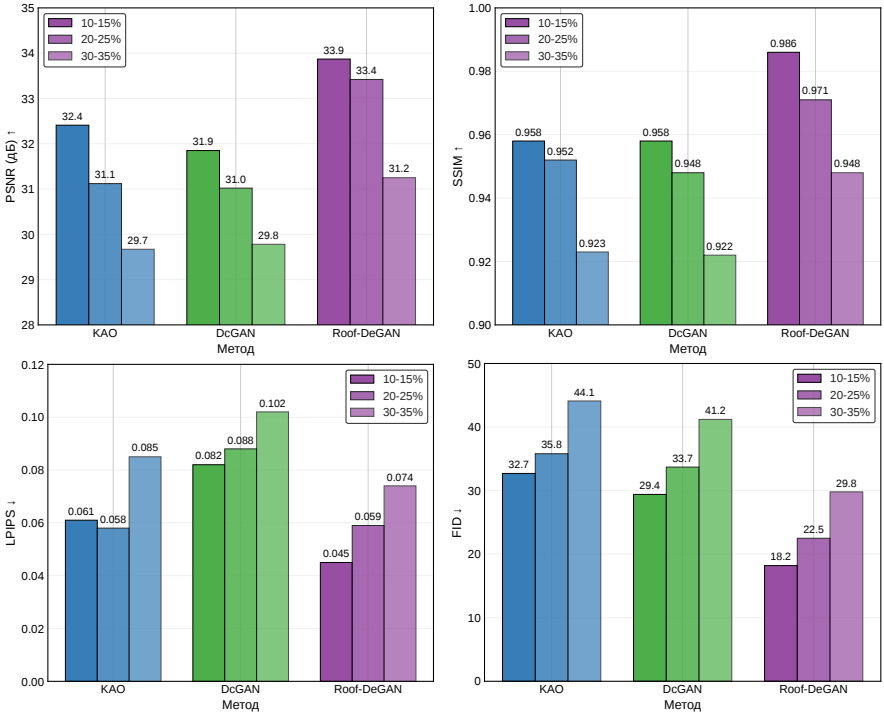


Рисунок 7. Сравнение геометрических (PSNR, SSIM) и перцептивных метрик (LPIPS, FID) Roof-DeGAN с лучшими конкурентами

В таблице 4 приведено сравнение методов по метрике качества границ Boundary F1. Красным выделен лучший результат.

Таблица 4. Сравнение методов по метрике качества границ Boundary F1 (усреднение по всем типам повреждений)

Методы		Boundary F1↑
Традиционные методы	Navier-Stokes [3]	0,38
	PatchMatch [4]	0,46
Диффузионные модели	DDPM [8]	0,62
	SatDiff [10]	0,79
	KAO [11]	0,82
Гибридные и GAN-методы	Pix2Pix [6]	0,54
	ESRGAN [7]	0,58
	DeGAN базовый [13]	0,81
Новый метод	Roof-DeGAN	0,91

Результаты успешного восстановления крыш на тестовом датасете ППК «Роскадастр» показаны на рисунке 8. На нём для каждого из

(*а*) Исходные изображения

(*б*) Результаты успешного восстановления

Рисунок 8. Крыши дачных домиков до и после их успешного восстановления с использованием предлагаемого метода

трёх примеров приведены исходное изображение (частично закрытое кронами) и результат работы предлагаемой модели. Степень сложности восстановления возрастает слева направо: 10% перекрытия (многоскатная крыша) –минимальное восстановление; 20% перекрытия (двускатная крыша) –восстановление текстуры и геометрии скатов; 35% перекрытия (многоскатная крыша) –значительная реконструкция углов и текстуры.

Для малых и средних повреждений (10–20%) качество восстановления является более чем приемлемым: геометрия скатов сохраняется, текстура кровельного материала воспроизводится достоверно. При обширных повреждениях (35% и более) на восстановленном изображении могут наблюдаться значительные искажения текстур кровельного покрытия и формы крыши. При этом доля полностью неудачных результатов не превышает 3–5% от общего числа обработанных снимков; в этих случаях качество остаётся неприемлемым для последующей автоматизированной кадастровой обработки, рисунок 9.

(*а*) Исходное изображение (*б*) Результаты неуспешного восстановления

Рисунок 9. Пример неуспешного восстановления изображения крыши с использованием предлагаемого метода

8. Ограничения метода и перспективы дальнейших исследований

Несмотря на высокие количественные и качественные результаты, модель Roof-DeGAN обладает рядом ограничений, открывающих направления для дальнейшего развития.

Зависимость от качества маски. Модель требует бинарной маски повреждённой области M. В реальных условиях автоматическая сегментация окклюзий (крон деревьев, теней, техники) редко достигает идеального

качества. Качественный анализ показывает, что при незначительных ошибках сегментации ($\text{IoU} \in [0{,}85; 0{,}95]$) модель сохраняет высокую устойчивость – артефакты локализуются преимущественно на границах маски и не затрагивают внутреннюю геометрию скатов благодаря skip-соединениям и многоуровневому дискриминатору. При существенных искажениях маски ($\text{IoU} < 0{,}8$) наблюдается деградация качества восстановления: появляются эффекты «двойных контуров», локальные искажения текстур и нарушение геометрии коньков, поскольку генератор либо пытается восстановить уже видимые участки, либо оставляет часть окклюзии нетронутой.

Разрешение изображений. Для обработки аэрофотоснимков высокого разрешения, значительно превышающих размер 256×256 пикселей, перспективным направлением является интеграция предложенной архитектуры Roof-DeGAN с фреймворками фрагментарного вывода, такими как SAHI (Slicing Aided Hyper Inference). Данный подход предполагает разбиение исходного изображения и соответствующей маски повреждения M на перекрывающиеся патчи фиксированного разрешения, независимую обработку каждого патча обученной моделью и последующую агрегацию результатов с применением взвешенного усреднения в зонах перекрытия для минимизации граничных артефактов. Такая стратегия позволяет сохранить вычислительную эффективность инференса при работе с изображениями размером 1024×1024 и более, обеспечить непрерывность текстур кровельных материалов и геометрическую согласованность архитектурных элементов крыши. В качестве альтернативного пути масштабирования может быть рассмотрен переход на более эффективные трансформерные блоки с линейной сложностью внимания (Swin Transformer v2 [**30**], EfficientViT [**31**] и др.).

Представленность редких классов и сценариев. Датасет охватывает основные типы кровельных покрытий (черепица, металлочерепица, шифер, битумная черепица), однако недостаточно представляет редкие материалы (солома, медь, сланец, зелёная кровля, мембранные покрытия) и сложные погодные условия (снег, дождь, тени от соседних зданий). Расширение датасета за счёт синтетических изображений и применение методов доменной адаптации позволит существенно повысить обобщающую способность модели.

Вычислительная сложность обучения. На этапе инференса модель демонстрирует высокую эффективность (0,15 с на изображение 256×256 на GPU NVIDIA Tesla A100 SXM). Однако обучение требует около 6 часов

машинного времени на том же оборудовании. Значительная ресурсоёмкость обучения ограничивает масштабирование на большие датасеты и изображения высокого разрешения. Перспективными направлениями оптимизации являются дистилляция знаний, квантование весов и замена базовой архитектуры на более лёгкие свёрточные сети [28].

Ограничения диффузионного усиления. Диффузионная компонента повышает стабильность генерации и реалистичность текстур, однако увеличивает вычислительную нагрузку и в отдельных случаях приводит к лёгкому сглаживанию мелких деталей при очень плотных масках. Дальнейшая оптимизация (работа в латентном пространстве или сокращение числа шагов диффузии) позволит устранить данный недостаток.

Чувствительность к локализации повреждений. Хотя модель демонстрирует высокую среднюю точность восстановления границ по метрике Boundary F1 (таблица 4), этот результат достигается в условиях, когда маски повреждений преимущественно расположены в центральных областях скатов (что соответствует 85% тестовой выборки). При обширных повреждениях (30–35%), затрагивающих границы крыш, точность падает до 0,68–0,74 (рисунок 9). Данное ограничение связано с тем, что skip-соединения и межмасштабное внимание не могут передать геометрию, если повреждена вся граница. Перспективным решением является интеграция 3D-каркасов крыш из датасета ZRG в процесс обучения.

Ограниченная применимость. Экспериментальная валидация модели проводилась исключительно на данных ППК «Роскадастр». Полученные результаты могут не обобщаться на аэрофотоснимки, полученные в иных условиях. Для расширения области применения модели требуется дополнительная валидация на других датасетах и, при необходимости, дообучение модели.

В текущей версии исследования целевой датасет ограничен 2000 изображениями. Для дальнейшего повышения обобщающей способности модели планируется расширение выборки до 5000+ изображений с включением редких типов кровельных покрытий и естественных окклюзий (облачность, сезонные изменения растительности). Перспективами развития также являются использование временных рядов аэрофотоснимков (восстановление по нескольким датам) и переход к трёхмерному восстановлению геометрии крыш.

9. Заключение

В ходе выполнения работы разработана гибридная генеративная модель Roof-DeGAN для восстановления скрытых областей крыш зданий на аэрофотоснимках. Предложенная архитектура сочетает трансформерные блоки для учёта глобального контекста, плотные свёрточные связи для улучшения распространения признаков и механизм межмасштабного внимания в многоуровневом дискриминаторе для повышения стабильности обучения.

Основные результаты работы:

- Разработана архитектура генератора типа энкодер–декодер с трансформерными блоками с динамическим разрежением внимания, снижающим сложность с $O(n^2)$ до $O(n \cdot k)$ за счёт адаптивного пропуска однородных областей снимка.
- Создан многоуровневый дискриминатор, оценивающий правдоподобность восстановленных фрагментов на разных масштабах, что повышает устойчивость обучения и качество текстур.
- Предложен и экспериментально обоснован двухэтапный метод обучения: предобучение на датасете ZRG в режиме сегментации крыш с последующим переносом весов в задачу восстановления. Показано, что такой подход обеспечивает прирост PSNR на 6,16 дБ по сравнению с обучением с нуля (таблица 2).
- Экспериментально подтверждена высокая эффективность предложенного подхода: на тестовой выборке достигнуты PSNR = 33,7 дБ, SSIM = 0,971, LPIPS = 0,048 и FID = 17,8, что превосходит современные методы на датасете ППК «Роскадастр» (таблица 1). На тестовой выборке средняя метрика Boundary F1 составила 0,91. При этом для повреждений площадью 10–15% значение достигает 0,96, для 20–25% –0,88, а для 30–35% снижается до 0,74 (таблица 4). Это подтверждает, что модель уверенно восстанавливает геометрию крыш при умеренных повреждениях, однако при обширных окклюзиях точность границ ожидаемо падает.

Полученные результаты могут быть использованы в системах автоматизированной обработки данных дистанционного зондирования Земли, при обновлении картографических материалов, в задачах мониторинга городской застройки и состояния кровель зданий, а также в смежных областях, требующих восстановления скрытых фрагментов изображений.

Дальнейшие исследования будут направлены на адаптацию разработанной модели для учёта временной динамики изменения растительности, интеграцию с данными других спектральных диапазонов, а также на применение предложенного подхода для решения смежных задач: устранения теней, восстановления повреждённых архивных снимков и улучшения качества изображений, полученных в неблагоприятных метеоусловиях.

Список использованных источников

[1] S. May, Y. Wang, L. Zhang *Building damage assessment with deep learning* // ISPRS Archives.– 2022.– Vol. **XLIII-B3-2022**.– Pp. 1133–1138. ↑228, 241

[2] L. Dong, J. Shan *A comprehensive review of earthquake-induced building damage detection with remote sensing techniques* // ISPRS Journal of Photogrammetry and Remote Sensing.– 2013.– Vol. **84**.– Pp. 85–99. ↑228

[3] M. Bertalmio, G. Sapiro, V. Caselles, C. Ballester *Image inpainting* // *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH 2000 (New Orleans, LA, USA, July 23–28, 2000).– ACM.– 2000.– ISBN 1-58113-208-5.– Pp. 417–424. ↑228, 245, 252, 254

[4] C. Barnes, E. Shechtman, A. Finkelstein, D. B. Goldman *PatchMatch: a randomized correspondence algorithm for structural image editing* // ACM Transactions on Graphics.– 2009.– Vol. **28**.– No. 3.– id. 24.– 11 pp. ↑228, 245, 252, 254

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio *Generative adversarial networks* // Communications of the ACM.– 2020.– Vol. **63**.– No. 11.– Pp. 139–144. ↑228, 234

[6] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros *Image-to-image translation with conditional adversarial networks* // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR 2017 (Honolulu, HI, USA, July 21–26, 2017).– IEEE.– 2017.– ISBN 978-1-5386-0457-1.– Pp. 5967–5976. ↑228, 234, 245, 252, 254

[7] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. C. Loy *ESRGAN: Enhanced super-resolution generative adversarial networks* // *Computer Vision – ECCV 2018 Workshops*, Proceedings.– V. V (Münich, Germany, September 8–14, 2018), Lecture Notes in Computer Science.– vol. **11133**.– Springer.– 2019.– ISBN 978-3-030-11020-8.– Pp. 63–79. ↑228, 245, 252, 254

[8] J. Ho, A. Jain, P. Abbeel *Denoising diffusion probabilistic models* // *Advances in Neural Information Processing Systems 33*, 34th Conference on Neural Information Processing Systems (NeurIPS 2020) (virtual, December 6–12, 2020).– 2020.– ISBN 9781713829546.– Pp. 6840–6851. arXiv 2006.11239 ↑229, 235, 245, 252, 254

[9] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, M. Norouzi *Image super-resolution via iterative refinement* // IEEE Transactions on Pattern Analysis and Machine Intelligence.– 2023.– Vol. **45**.– No. 4.– Pp. 4713–4726. doi ↑229, 235

[10] T. Panboonyuen, C. Charoenphon, C. Satirapod *SatDiff: A stable diffusion framework for inpainting very high-resolution satellite imagery* // IEEE Access.– 2025.– Vol. **13**.– Pp. 51617–51631. doi ↑229, 245, 252, 254

[11] T. Panboonyuen *KAO: Kernel-adaptive optimization in diffusion for satellite image* // IEEE Transactions on Geoscience and Remote Sensing.– 2025.– Vol. **63**.– id. 5531217.– 17 pp. doi ↑229, 245, 252, 254

[12] Y. Zhou, X. Gao, X. Wu, F. Wang, W. Jing, X. Hu *Image characteristic-guided learning method for remote-sensing image inpainting* // Remote Sensing.– 2025.– Vol. **17**.– No. 13.– id. 2132.– 22 pp. doi ↑229, 235, 236

[13] R. Li, L. Wen, S. Shao, M. Yu, L. Mohaisen *A novel generative adversarial network framework for super-resolution reconstruction of remote sensing* // Frontiers in Earth Science.– 2025.– Vol. **13**.– id. 578321.– 17 pp. doi ↑229, 245, 252, 254

[14] Z. Zhang, W. Feng, M. Zhong, M. Yang *BD-VITGAN: A blind dense VITGAN for satellite remote sensing images super-resolution reconstruction* // Geo-spatial Information Science.– 2025.– Pp. 1–23. doi ↑229

[15] Y. Wang, W. Wu, Z. Zhang, Z. Li, F. Zhang, X. Li *A temporal attention-based multi-scale generative adversarial network to fill gaps in time series of MODIS data for land surface phenology extraction* // Remote Sensing of Environment.– 2025.– Vol. **318**.– id. 114507. doi ↑229

[16] D. Zhou, L. Xu, K. Wu, H. Liu, M. Jiang *DSEPGAN: A dual-stream enhanced pyramid based on generative adversarial network for spatiotemporal image fusion* // Remote Sensing.– 2025.– Vol. **17**.– No. 24.– id. 4050.– 25 pp. doi ↑229

[17] И. В. Винокуров *Повышение точности сегментирования объектов с использованием генеративно-состязательной сети* // Программные системы: теория и приложения.– 2025.– Т. **16**.– № 2.– С. 111–152 (Англ., Рус.). doi ↑229

[18] И. В. Винокуров *Использование модели Mask R-CNN для сегментации объектов недвижимости на аэрофотоснимках* // Программные системы: теория и приложения.– 2025.– Т. **16**.– № 1.– С. 3–44 (Англ., Рус.). doi ↑229

[19] J. Johnson, A. Alahi, L. Fei-Fei *Perceptual losses for real-time style transfer and super-resolution* // *Computer Vision - ECCV 2016*, Proceedings.– V. II, 14th European Conference (Amsterdam, The Netherlands, October 11–14, 2016), Lecture Notes in Computer Science.– vol. **9906**.– Springer.– 2016.– ISBN 978-3-319-46474-9.– Pp. 694–711. doi ↑235, 237

[20] J. Zhang, Y. Xiao, G. Chen, Q. Sun, F. Xu, C.-S. Leung *Histogram-guided semantic-aware colorization* // *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 2022 (Virtual and Singapore, May 23–27, 2022).– IEEE.– 2022.– ISBN 978-1-6654-0541-6.– Pp. 2549–2553. doi ↑235

[21] I. Corley, J. Lwowski, P. Najafirad *ZRG: A dataset for multimodal 3D residential rooftop understanding* // *2024 IEEE/CVF Winter Conference on Applications of Computer Vision*, WACV 2024 (Waikoloa, HI, USA, January 03–08, 2024).– IEEE.– 2024.– ISBN 979-8-3503-1893-7.– Pp. 4623–4631. doi arXiv 2304.13219 % doi ↑236

[22] E. Maggiori, Y. Tarabalka, G. Charpiat, P. Alliez *Can semantic labeling methods generalize to any city? The INRIA aerial image labeling benchmark* // *2017 IEEE International Geoscience and Remote Sensing Symposium*, IGARSS 2017 (Fort Worth, TX, USA, July 23–28, 2017).– IEEE.– 2017.– ISBN 978-1-5090-4951-6.– Pp. 3226–3229. doi ↑236

[23] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Bailard, S. Benitez, U. Breitkopf *The ISPRS benchmark on urban object classification and 3D building reconstruction* // ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences.– 2012.– Т. **I**.– № 3.– С. 293–298. doi URL ↑236

[24] Q. Huynh-Thu, M. Ghanbari *Scope of validity of PSNR in image/video quality assessment* // Electronics Letters.– 2008.– Vol. **44**.– No. 13.– Pp. 800–801. doi ↑241

[25] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *Image quality assessment: from error visibility to structural similarity* // IEEE Transactions on Image Processing.– 2004.– Vol. **13**.– No. 4.– Pp. 600–612. doi ↑241

[26] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang *The unreasonable effectiveness of deep features as a perceptual metric* // *2018 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2018 (Salt Lake City, UT, USA, June 18–22, 2018).– IEEE.– 2018.– ISBN 978-1-5386-6421-6.– Pp. 586–595. doi ↑241

[27] A. Sekrecka, K. Karwowska *Classical vs. machine learning-based inpainting for enhanced classification of remote sensing image* // Remote Sensing.– 2025.– Vol. **17**.– No. 7.– id. 1305.– 36 pp. doi ↑241

[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter *GANs trained by a two time-scale update rule converge to a local Nash equilibrium* // *Advances in Neural Information Processing Systems 30*, 31st Annual Conference on Neural Information Processing Systems 2017 (Long Beach, CA, USA, December 4–9, 2017).– 2017.– ISBN 9781510860964.– Pp. 6626–6637. doi URL ↑241, 257

[29] D. Hasler, S. E. Süsstrunk *Measuring colourfulness in natural images*, SPIE/IS&T Human Vision and Electronic Imaging (Santa Clara, CA, United States, 20 January 2003), Proceedings of SPIE.– vol. **5007**.– 2003.– Pp. 87–95. doi ↑251

[30] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, B. Guo *Swin Transformer V2: Scaling up capacity and resolution* // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR 2022 (New Orleans, LA, USA, June 18–24, 2022).– IEEE.– 2022.– ISBN 978-1-6654-6946-3.– Pp. 11999–12009. doi ↑256

[31] H. Cai, J. Li, M. Hu, C. Gan, S. Han *EfficientViT: Multi-scale linear attention for high-resolution dense prediction.*– 2024.– 12 pp. arXiv 2205.14756 ↑256

Поступила в редакцию 30.04.2026;
одобрена после рецензирования 01.06.2026;
принята к публикации 10.06.2026;
опубликована онлайн 20.06.2026.

Рекомендовал к публикации *к.т.н. В.П. Фраленко*

Информация об авторах:

Игорь Викторович Винокуров

Кандидат технических наук (PhD), ассоциированный профессор в Финансовом Университете при Правительстве Российской Федерации. Область научных интересов: информационные системы, информационные технологии, технологии обработки данных

0000-0001-8697-1032
e-mail: *igvvinokurov@fa.ru*

Георгий Михайлович Лапаньков

Выпускник (бакалавр) Финансового Университета при Правительстве Российской Федерации. Область научных интересов: информационные системы, разработка мобильных приложений, анализ данных

0009-0007-0511-628X
e-mail: *goshmen2004@gmail.com*

Георгий Дмитриевич Умаров

Выпускник (бакалавр) Финансового Университета при Правительстве Российской Федерации. Область научных интересов: информационные технологии, веб-разработка, анализ данных

0009-0007-0364-8477
e-mail: *goshaumarov0609@mail.ru*

Вклад авторов: *И.В. Винокуров* – 70% (разработка модели и методики проведения экспериментов, реализация обучения и исследования модели, интеграция результатов в информационные системы ППК «Роскадастр»); *Г.М. Лапаньков* – 15% (реализация предобучения на ZRG); *Г.Д. Умаров* – 15% (формирование синтетических масок, визуализация результатов обучения модели).

Декларация об отсутствии личной заинтересованности: *благополучие авторов не зависит от результатов исследования.*